# LAW
# And the Philosophy of Action

*Social, Political & Legal Philosophy, Volume 3*

Edited by
Enrique Villanueva

INSTITUTO DE INVESTIGACIONES JURÍDICAS
UNIVERSIDAD NACIONAL AUTONOMA DE MÉXICO

SOCIAL, POLITICAL, AND LEGAL PHILOSOPHY



The paper on which this book is printed meets the requirements of "ISO 9706:1994, Information and documentation - Paper for documents - Requirements for permanence".

ISBN: 978-90-420-3853-0
E-Book ISBN: 978-94-012-1096-6

Printed in the Netherlands

## Contents

Social, Political, & Legal Philosophy, 3
Law: and the Philosophy of Action, 2014

# What Temptation Could Not Be: A Lesson from the Criminal Law

Gabriel S. Mendlow*

**Abstract.** Prominent theories of the criminal law borrow heavily from the two leading theories of temptation—the evaluative conception of temptation, which conceives emotion and desire as essentially involving a kind of evaluation, and the mechanistic conception of temptation, which conceives emotion and desire as essentially involving felt motivation. As I explain, both conceptions of temptation are inconsistent with the possibility of akratic action, that is, action contrary to a person's conscious better judgment. Both are inconsistent with the possibility of akratic action because both are covertly inconsistent with a two-fold psychological assumption that undergirds common beliefs about human action and lies at the heart of the law of criminal responsibility: that resisting a powerful temptation is extremely difficult yet not ordinarily impossible. I reveal these inconsistencies and offer in place of the leading theories of temptation a theory of affective desire as primitive psychic attraction, an elemental psychological state typically accompanied by evaluation and motivation but not reducible to either one. I then show how this theory of desire is consistent with the possibility of akratic action, with the two-fold psychological assumption at the heart of the law of criminal responsibility, and, in particular, with the defense of provocation.

* Assistant Professor of Law and Assistant Professor of Philosophy University of Michigan

## I. Introduction

The law sometimes excuses criminal misconduct that results from a person's failure to resist a powerful temptation. In certain cases, the temptation is a pathological urge—the "irresistible" impulse of one who invokes the defense of insanity,[1] for example, or the chemical compulsion of one who invokes the defense of involuntary intoxication.[2] In other cases, the temptation is but an extreme version of something mundane and familiar. Here we meet with the defenses of duress[3] and provocation,[4] defenses that apply to defendants who succumb to a temptation that is at once all too powerful and all too understandable—

[1] In many jurisdictions, the traditional *M'Naghten* rule for insanity is supplemented by an "irresistible impulse" test, which "requires a verdict of not guilty by reason of insanity if it is found that the defendant had a mental disease which kept him from controlling his conduct." 1 Wayne R. LaFave, Substantive Criminal Law 545 (2003). The "irresistible impulse" test originated in an English case of 1840, in which the defendant was charged with treason for firing a pistol at the Queen (he missed). The judge instructed the jury that "[i]f some controlling disease was, in truth, the acting power within [the defendant] which he could not resist, then he will not be responsible." 1 LaFave, supra, at 545.

[2] Jurisdictions that permit the defense of involuntary intoxication typically formulate the defense in parallel to some formulation of the insanity defense, such that "[a]n actor is excused for his conduct constituting an offense if, as a result of . . . [involuntary] intoxication, ...the actor (a) does not perceive the physical nature or consequences of his conduct, or (b) does not know his conduct is wrong or immoral, or (c) is not sufficiently able to control his conduct so as to be held accountable for it." 2 Paul H. Robinson, Criminal Law Defenses 339 (1984).

[3] LaFave explains the traditional defense of duress as follows: "A person's unlawful threat (1) which causes the defendant reasonably to believe that the only way to avoid imminent death or serious bodily injury to himself or to another is to engage in conduct which violates the literal terms of the criminal law, and (2) which causes the defendant to engage in that conduct, gives the defendant the defense of duress (sometimes called compulsion or coercion) to the crime in question unless the crime consists of intentionally killing an innocent third person." 2 LaFave, supra note 1, at 72.

[4] The common-law doctrine of provocation downgrades an intentional homicide from murder to voluntary manslaughter in the case of a defendant who killed the victim upon "adequate provocation" in the "sudden heat of passion" and in the absence of sufficient "cooling time." 2 LaFave, supra note 1, at 775-78. Section 210.3(1)(b) of the Model Penal Code (1980) formulates the doctrine of provocation more broadly, stating that "[c]riminal homicide constitutes manslaughter when . . . a homicide which would otherwise be murder is committed under the influence of extreme mental or emotional disturbance for which there is reasonable explanation or excuse."

in the case of duress, the temptation to yield to a violent threat; in the case of provocation, the temptation to retaliate against an outrageous affront.

Common to all of these volition-based defenses is a crucial psychological assumption: that resisting a powerful impulse is extremely difficult. Common specifically to provocation and duress is a further assumption: that no matter how difficult it might be to resist a powerful impulse, doing so is not ordinarily impossible. This further assumption helps explain why the defenses of provocation and duress are *incomplete* defenses, in that provocation mitigates but does not fully exculpate, and duress fully exculpates but does not apply to all crimes.

Each of these psychological assumptions follows in turn from a simple but compelling truism about the nature of temptation. The first assumption follows from the truism that resisting a temptation is ordinarily difficult; the second, from the truism that resisting a temptation is not ordinarily impossible. (Before going any further, I should clarify that, by a *temptation*, I mean any psychological state—whether an emotion or a feeling or a desire—that can lead a person to act *akratically*, that is, to act contrary to her conscious better judgment.[5]) Basic to our commonsense conception of human action, these two truisms about temptation undergird and explain our practices of blame and sanction. That resisting a temptation is ordinarily difficult (the first truism) explains why volitional impairment typically reduces a wrongdoer's blameworthiness. That resisting a temptation is not ordinarily impossible (the second truism) explains why volitional impairment rarely reduces a wrongdoer's blameworthiness to nothing.

The trouble, I will argue, is this: each truism conflicts with one of the two conceptions of temptation that have long dominated Western philosophy. The first truism—that resisting a temptation is ordinarily difficult—conflicts with the *evaluative conception* of temptation: if temptation consists in the 'appearance' of values or reasons, as the evaluative conception supposes, then resistance will turn out to be too

[5] The root of the words *akratic* and *akratically* is the ancient Greek word *akrasia*, which literally means weakness. For a discussion of the various forms of akratic action, see Gabriel S. Mendlow, Want of Care: An Essay on Wayward Action, Ethical Theory and Moral Practice, DOI: 10.1007/s10677-013-9436-1 (2013).

easy—indeed, it will turn out to be effortless. The second truism—that resisting temptation is not ordinarily impossible—conflicts with the *mechanistic conception* of temptation: if temptation consists in felt motivation, as the mechanistic conception supposes, then resistance will turn out to be too hard—indeed, it will turn out to be impossible. Neither of these conflicts is anything close to obvious, or else there would be little need for this essay.

The upshot is that the prevailing theories of temptation are incompatible with the criminal law's volition-based defenses because they are incompatible with the two truisms about temptation that those defenses presuppose. Furthermore, precisely because the prevailing theories of temptation are incompatible with the two truisms, these theories fail to make sense of how people can act akratically.[6] The prevailing theories of temptation actually make action contrary to one's better judgment seem impossible. As the essay will explain, the mechanistic conception ultimately (if non-obviously) entails that no one ever acts contrary to her better judgment voluntarily, and the evaluative conception ultimately (if also non-obviously) entails that no one ever acts contrary to her better judgment at all, whether voluntarily or otherwise.

Few proponents of the leading theories of temptation would embrace these conflicts outright and deny the possibility of akratic action or reject one of the two truisms about temptation that akratic action seemingly exemplifies. Akratic action is widely thought possible and the two truisms are central to our practices of blame and sanction. The truisms are also deeply intuitive. No one wants to give them up, and no one should. My primary aim, therefore, is to show that a strict proponent of either theory of temptation will have no choice but to give up at least one of the truisms and thereby reject the moral foundation of a central part of the criminal law. Achieving this aim will have two immediate consequences. The first will be to undermine theories of the

[6] Although philosophers occasionally have offered arguments to show that akratic action is impossible, see Plato, Protagoras 358b-c; R.M. Hare, The Language of Morals (1952); R.M. Hare, Freedom and Reason (1963); none of these arguments commands anything close to widespread assent. Nor is there any empirical evidence indicating that people never act akratically. (There is some empirical evidence purportedly suggesting that akratic action is not as common as many of us believe. See Richard Holton, Willing, Wanting, Waiting 97-111 (2009).)

criminal law that employ either of the leading conceptions of temptation. The second will be to undermine these conceptions of temptation themselves.

After arguing against the prevailing philosophical conceptions of temptation, I will sketch an alternative theory of one important variety of temptation—affective desire—and show how this theory is consistent both with commonsense psychology and with the criminal law, focusing for demonstrative purposes on the defense of provocation.

## II. The Evaluative Conception of Temptation

The evaluative conception of temptation is dominant among theories of emotion and desire.[7] And for good reason: evaluative theories of emo-

[7] For evaluative theories of emotion, see, e.g., Plato, The Republic; Aristotle, On the Soul; 2 The Hellenistic Philosophers 404-18 (A.A. Long & D.N. Sedley eds., 1987) (on the Stoics); C.D. Broad, Emotion and Sentiment, *in* Broad, Critical Essays in Moral Theory (1971); William Lyons, Emotion (1980); Robert C. Solomon, The Passions: Emotions and the Meaning of Life (1993); Jerome Neu, A Tear is an Intellectual Thing (2000); Christine Tappolet, Emotions et Valeurs (2000); Martha C. Nussbaum, Upheavals of Thought: The Intelligence of Emotions (2001); Jesse J. Prinz, Gut Reactions: A Perceptual Theory of Emotion (2004). See generally Dan M. Kahan and Martha C. Nussbaum, Two Conceptions of Emotion in Criminal Law, 96 Colum. L. Rev. 269 (1996), at 289-93, for a concise history of the evaluative conception of emotion, replete with references to historical figures and contemporary theorists. For evaluative theories of desire, see, e.g., Aristotle, On the Soul 433a27-29, *in* The Complete Works of Aristotle: The Revised Oxford Translation (Jonathan Barnes ed., 1984); Donald Davidson, How Is Weakness of the Will Possible?, *in* Davidson, Essays on Actions and Events (1970); Dennis W. Stampe, The Authority of Desire, 96 Phil. Rev. 335 (1987); T.M. Scanlon, What We Owe to Each Other (1998); T.M. Scanlon, Reasons and Passions, *in* Contours of Agency: Essays on Themes from Harry Frankfurt (Sarah Buss and Lee Overton eds., 2002); R. Jay Wallace, Addiction as Defect of the Will: Some Philosophical Reflections, *in* Wallace, Normativity and the Will (2007); R. Jay Wallace, Three Conceptions of Rational Agency, *in* Wallace, supra; R. Jay Wallace, Normativity, Commitment, and Instrumental Reason, *in* Wallace, supra; Mark Johnston, The Authority of Affect, 63 Philosophy and Phenomenological Research 181 (2001); Sergio Tenenbaum, The Judgment of a Weak Will, 59 Philosophy and Phenomenological Research 875 (1999); Sergio Tenenbaum, *Accidie*, Evaluation, and Motivation, *in* Weakness of Will and Practical Irrationality (Sarah Stroud and Christine Tappolet eds., 2003); Sergio Tenenbaum, Appearances of the Good (2007); Graham Oddie, Value, Reality, and Desire (2005); Jennifer S. Hawkins, Desiring the Bad Under the Guise of the Good, 58 Phil. Quarterly 244 (2008).

tion and desire offer penetrating accounts of the character, content, and moral significance of the various psychological states that define us as creatures who *feel* as well as think. I will argue that these theories nevertheless fail to account for the particular phenomenon of volitional impairment—the difficulty of hewing to one's better judgment in the face of temptation—and for this reason fail as theories of the criminal law.

As Dan Kahan and Martha Nussbaum describe the evaluative conception of emotion,

> the emotions themselves *contain* an evaluation or appraisal of [their objects]... Grief sees the lost one as of enormous significance; so too, in a happier way, does love. Disgust usually sees the object as one that threatens or contaminates, one that needs to be kept at a distance from the self. Fear perceives the impending harm as significant; anger sees the wrong as pretty large—whether or not this is the way these things really are.[8]

Along similar lines, the evaluative conception of desire conceives a desire as a quasi-cognitive (or perhaps quasi-perceptual) way of regarding the desired object as good or valuable. To desire something, according to this view, is to see the thing as good or valuable, or to see the thing as something that there is reason to obtain or bring about. Common to the evaluative conceptions of emotion and desire is the idea that temptation's essence is evaluative appearance: temptation makes its object *appear* good or *look* good or *seem* good. It is a further question whether the subject of temptation actually *believes* the tempting object to be good. An object can appear good to us without our really believing that it is—or so the evaluative conception of temptation maintains.[9]

[8] Kahan and Nussbaum, Two Conceptions of Emotion, supra note 7, at 285.

[9] It is important to distinguish between two claims that sound similar but in fact are independent. The first claim, which I have been calling the evaluative conception of desire, is that desiring something is a way of regarding it as good. The second claim is that we can desire something *intelligibly* only if we find it good in some respect and to some degree; see G.E.M. Anscombe, Intention (1957); Joseph Raz, Incommensurability and Agency, *in* Raz, Engaging Reason: On the Theory of Value and Action (1999); Joseph Raz, Agency, Reason, and the Good, *in* Raz, supra. The second claim

Besides being important to philosophers, the evaluative conception of temptation is influential with theorists of the criminal law.[10] Most prominently, Kahan and Nussbaum see the evaluative conception as undergirding everything from the voluntary act requirement and the doctrine of premeditated murder to the defenses of insanity, self-defense, duress, and provocation.[11] Illustrative of the evaluative approach to these doctrines is Kahan and Nussbaum's treatment of the provocation defense, a defense that reduces murder to voluntary manslaughter in the case of a defendant who killed the victim upon "adequate" provocation in the "heat of passion" and in the absence of sufficient "cooling time.."[12]

The evaluative theory of the provocation defense is most easily understood in contrast to the leading alternative,[13] the *volitional theory*.[14] The volitional theory roots the mitigating effect of provocation in volitional impairment: a hot-blooded killer is less blameworthy than a cold-blooded one because the hot-blooded killer has far more trouble

is not a conception of desire so much as a constraint on desire—a constraint that is wholly consistent with the argument of this essay.

[10] See, e.g., Kahan and Nussbaum, Two Conceptions of Emotion, supra note 7; Samuel H. Pillsbury, Emotional Justice: Moralizing the Passions of Criminal Punishment, 74 Cornell L. Rev. 655, 678 (1989).

[11] Kahan and Nussbaum, Two Conceptions of Emotion, supra note 7.

[12] See 2 LaFave, supra note 1, at 775-78.

[13] A third approach explains the provocation defense on utilitarian grounds, holding that hot-blooded killers should be punished less than cold-blooded ones either because the former are either less dangerous or less able to be deterred. See, e.g., Jeremy Bentham, An Introduction to the Principles of Morals and Legislation; Jerome Michael & Herbert Wechsler, A Rationale of the Law of Homicide II, 37 Colum. L. Rev. 1261 (1937); Gary S. Becker, Crime and Punishment: An Economic Approach, 76 J. Pol. Econ. 169 (1968); Richard A. Posner, An Economic Theory of the Criminal Law, 85 Colum. L. Rev. 1193 (1985); Steven Shavell, Criminal Law and the Optimal Use of Nonmonetary Sanctions as a Deterrent, 85 Colum. L. Rev. 1232 (1985). The utilitarian rationale for the provocation defense has declined in recent decades as retributivism has become the reigning penal philosophy.

[14] For volitional accounts of the provocation defense, see, e.g., George P. Fletcher, Rethinking Criminal Law 242-43 (1978); Joshua Dressler, Rethinking Heat of Passion: A Defense in Search of a Rationale, 73 J. Crim. L. & Criminology 421 (1982); Joshua Dressler, When "Heterosexual" Men Kill "Homosexual" Men: Reflections on Provocation Law, Sexual Advances, and the "Reasonable Man" Standard, 85 J. Crim. L. & Criminology 726 (1995); Joshua Dressler, Why Keep the Provocation Defense?: Some Reflections on a Difficult Subject, 86 Minn. L. Rev. 959 (2002).

exercising self-control. In the words of Joshua Dressler, the leading volitional theorist,

> [t]he true reason for the law's "concession to human weakness"[15]—the reason why, if *A* kills *P* in sudden rage at his actions, the law will likely allow *A* to argue that the jury should reduce the homicide to manslaughter—is that the homicide is the result of an understandable and *excusable* loss of self-control arising from his anger. Common experience teaches that, at some point, anger becomes so intense that people find it extremely difficult to control themselves and respond constructively, rather than violently, to the anger-producing stimulus. Therefore, when *A* kills *P* because his reason is "disturbed or obscured by passion to an extent which *might render* ordinary men, of fair average disposition, *liable* to act rashly or without due deliberation or reflection, and from passion, rather than judgment",[16] he is less to blame than if he killed *P* while he was calm. This is because it is harder for *A* to control his actions when he is angry than when he is calm.[17]

Because it roots mitigation in volitional impairment, the volitional theory interprets each element of the provocation defense—adequate provocation, heat of passion, absence of cooling time—in terms of loss of self-control. A provocation is *adequate*, according to the volitional theory, if it is sufficient to cause a person endowed with ordinary powers of self-restraint to have serious difficulty controlling herself;[18] a person acts in the *heat of passion* if she kills under the influence of an emotion so strong that it "dominates [her] volition";[19] and a person kills in the absence of *reasonable cooling time* if she acts before a person of ordinary self-restraint would have regained control, so that "the killing [can be] seen as an outgrowth of the provocative event, [instead

[15] 2 American Law Inst., Model Penal Code and Commentaries, § 210.3 commentary, at 55.

[16] Maher v. People, 10 Mich. 212, 220 (1862).

[17] Dressler, When "Heterosexual" Men Kill "Homosexual" Men, supra note 14, at 747-48 (some internal citations omitted).

[18] LaFave defines an adequate provocation as "a provocation which would cause a reasonable man to lose his normal self-control" (2 Wayne R. LaFave, Substantive Criminal Law 491 (2003)). Cf. Dressler, Rethinking Heat of Passion, supra note 14, at 466-67.

[19] Smith v. State, 3 So. 551, 552 (Ala. 1888).

of] as an independent act for which [she] is fully accountable."[20] The reason why provocation mitigates instead of fully exculpating, on this view, is that the provoked killer does not wholly lack the capacity to control herself: however difficult it might be for an "adequately" provoked person to refrain from acting on her violent passion, refraining is not altogether impossible.[21]

By contrast, the evaluative theory of the provocation defense shifts the focus away from volitional impairment and interprets each element of the doctrine as an indicator not of a passion that is especially intense but of a passion that is especially appropriate. According to the evaluative theory, a provocation is *adequate* not (necessarily) when it induces a passion that overcomes one's self-control but instead when it is an affront in response to which one rightly becomes impassioned. A defendant acts in the *heat of passion* not (necessarily) when her emotion is so strong that it "dominates [her] volition"[22] but when the defendant's emotion embodies "an appropriate valuation of the good... that is threatened by the victim's wrongful provocation."[23] And a person kills in the absence of *reasonable cooling time* when she acts while it is (still) morally appropriate to be impassioned. According to the evaluative theory, the reason why provocation mitigates instead of fully exculpating is that the defendant's passion embodies a reasonable but imperfect valuation of the good that is threatened by the victim's wrongful provocation: the defendant is right to value the threatened good, but wrong to value it so much more than she evidently values the victim's life.[24]

[20] Fletcher, Rethinking Criminal Law, supra note 14, at 244.

[21] See Dressler, Why Keep the Provocation Defense? supra note 14, at 974.

[22] *Smith*, 3 So. at 552.

[23] Kahan and Nussbaum, Two Conceptions of Emotion, supra note 7, at 315 (1996).

[24] "To make this concrete," Kahan and Nussbaum write, "imagine a woman who kills a man in anger after discovering that he has sexually abused the woman's young daughter. From the evaluative perspective, one would say that her emotion embodies appraisals of mixed quality. She has appraised her circumstances in a way correctly, since her anger reflects her appropriate valuation of her daughter's well-being; but in a way she has also appraised them wrongly, since she should not have thought that this good was all-important, taking precedence over all other considerations, including the value of the man's life and the importance of lawful resolution of disputes. Her judgment may have been distorted because she harbored a skewed relative valuation of

Against the volitional theory, evaluative theorists rely heavily on common-law judicial opinions that have insisted that the adequacy of a given provocation goes not to passion's intensity but to passion's moral quality. As the Michigan Supreme Court asserted in an oft-cited nineteenth century provocation case, a "provocation [cannot] be held sufficient or reasonable [simply] because... a state of excitement [i.e., an intense passion] has followed from it; for then, by habitual and long continued indulgence of evil passions, a bad man might acquire a claim to mitigation which would not be available to better men, and on account of that very wickedness of heart which, in itself, constitutes an aggravation both in morals and in law."[25] As another nineteenth century court explained, no matter how intense a defendant's passion might be, such passion will not lessen the defendant's culpability if it springs from "a heart... devoid of social duty, that is, reckless of the rights and lives of others, and fatally bent on taking life...".[26] That court went on to say that, for a killing to qualify as manslaughter rather than murder, the "killing [must] proceed [not] from a bad or corrupt heart, [but] rather from the infirmity of passion to which even good men are subject."[27]

Now, even if the evaluative theory truly fits better with the common-law doctrine of provocation than the volitional theory does, the former will not succeed as a theory of the provocation defense unless it can make psychological sense of how the temptation to retaliate against an outrageous affront actually leads us to engage in retaliatory action. In fact, if the evaluative theory cannot make sense of how temptation leads us astray, it will not succeed as a theory of *any* volition-based defense. Furthermore, to succeed specifically as a theory of the provocation defense, the evaluative theory must accommodate the

the different goods involved or, more subtly, because she focused so intensely on one of them that other relevant considerations were temporarily eclipsed from view. But either way, our assessment of her behavior is likely to be complex; her emotional motivation is reasonable, but imperfect. The mitigating consequence of voluntary manslaughter captures the complexity of this assessment." Kahan and Nussbaum, Two Conceptions of Emotion, supra note 7, at 313 (internal citations omitted).

[25] Maher v. People, 10 Mich. 212, 220 (1862) (cited by Kahan and Nussbaum, Two Conceptions of Emotion, supra note 7, at 307).

[26] State v. Cook, 3 Ohio Dec. Reprint 142, 147 (1859).

[27] *Id.* at 146 (cited by Kahan and Nussbaum, Two Conceptions of Emotion, supra note 7, at 307).

phenomenon of akratic action. That is because the prototypical case of provoked retaliation is one in which the defendant acts contrary to her better judgment. Rarely does a provoked defendant kill because she suddenly has changed her mind about whether homicide is morally and legally permissible.[28] We should not be misled by the fact that the provoked defendant typically believes that her provoker deserves to be harmed or killed. A person can believe fervently and passionately that her provoker deserves to be killed while at the same time believing that it would be wrong for her to be the one to kill him. Even in moments of intense anger, we rarely stop believing that it is wrong to take the law into our own hands. (Consider the difference between believing that a particular person should be executed and believing that that person should be executed *by you*.) But even when we *do* stop believing that it is wrong to take the law into our own hands—even when we do temporarily change our minds about the moral permissibility of revenge killing—we generally do not change our minds about whether it would be a good idea to perform an action that will very likely send us to prison for a long, long time. The point is this: even if you suppose that the provoked defendant suddenly and temporarily forms the belief that revenge killing is morally permissible, it strains credulity to suppose that she also forms the belief that it is worth going to prison for many years in order to perpetrate *this* particular revenge killing.

Evaluative theorists therefore must be able to explain how a temptation *qua* evaluative appearance can lead us to act akratically. As I will argue, they cannot. The reason they cannot, I will explain, is that they are unable to reconcile the evaluative theory with the first truism about temptation, that resisting temptation is ordinarily difficult. The evaluative theory consequently turns out to be inconsistent with the host of volition-based criminal defenses, all of which presuppose the first truism. If temptation consists in the 'appearance' of values or reasons, resistance to temptation will end up being implausibly easy, indeed, practically effortless.

The standard account of how temptation *qua* evaluative appearance leads us astray relies on an analogy between akratic action and

[28] Here I disagree with Stephen P. Garvey, who asserts that the provoked defendant is often someone who "violates the law …because he honestly and momentarily …believes the law allows him to kill" (Passion's Puzzle, 90 Iowa L. Rev. 1677, 1683 (2005)).

perceptual illusion.[29] As the evaluative theorist Christine Tappolet explains, akratic action involves

> a conflict between a value perception and an evaluative judgment that can be compared to perceptual illusions such as the Muller-Lyer illusion, in which one sees the lines as being of a different length even though one judges or even knows that they are of the same length.[30]

Tappolet illustrates this comparison with several examples:

> Suppose I am about to cross a narrow rope bridge hanging high up on a deep shaft. Though I feel fear, I judge that all things considered I ought to cross the bridge; I judge it to be sufficiently safe and going back would make for a much longer hike. If I end up not crossing the bridge, it will not be difficult to make sense of my action: the perceived danger, be it real or not, readily explains why I don't cross the bridge. Or consider... [Dante's] Francesca and her passionate love for Paolo... [T]his love consists in the perception of Paolo as a worthy object of love. Now this cannot directly make Francesca's action intelligible, for it does not involve a perception of the value of the action itself. But it is surely an important part of what makes it intelligible. The value that Francesca perceives makes her desire to make love to Paolo intelligible and thus indirectly makes her action intelligible... even though [she] judge[s] that another course of action would have been better all things considered.[31]

These examples are meant to show that temptation *qua* evaluative appearance causes us to act akratically by functioning as a sort of practical illusion—an illusion that presents certain considerations

[29] Evaluative accounts of akratic action appear in: Christine Tappolet, Emotions and the Intelligibility of Akratic Action, *in* Stroud and Tappolet, supra note 7, at 111; Scanlon, What We Owe, supra note 7, at 35; Wallace, Addiction, supra note 7, at 185-86; Johnston, supra note 7, at 214; Tenenbaum, *Accidie*, supra note 7; Talbot Brewer, The Character of Temptation: Towards a More Plausible Kantian Moral Psychology, 83 Pac. Phil. Quarterly 103 (2002). See also Jessica Moss, *Akrasia* and Perceptual Illusion, 91 Archiv f. Gesch. d. Philosophie 119 (2009), which argues that Aristotle's account of akratic action entails that akratic action "is parallel to certain cases of perceptual illusion" (119).

[30] Tappolet, Emotions and the Intelligibility of Akratic Action, supra note 29, at 111.

[31] *Id.*

as being weightier than we believe them to be. On this view, the provoked defendant succumbs to a practical illusion to the effect that retaliation is justified. He knows that retaliation is in fact unjustified or at least a very bad idea—he knows, in other words, that the practical illusion is an illusion—but he retaliates anyway, because the practical illusion is especially captivating.

The standard evaluative account has a certain theoretical elegance. Yet it ultimately fails to explain akratic action because it fails to accommodate the first truism about temptation—that resisting temptation is ordinarily difficult—in that it fails to make sense of how we could have any trouble resisting evaluative illusions *that we recognize as illusions*. If I really do believe that *Y* is better than *X*—as I must, if my pursuit of *X* is to be truly akratic—how can *X*'s seeming or appearing good prompt me all by itself to pursue *X*? As we saw, Tappolet and others answer this question by exploiting an analogy between, on the one hand, evaluative appearance and action and, on the other, perceptual illusion and belief. But this strategy suffers from a problem: we would quite likely reject the counterpart story about perceptual illusion and belief.

Suppose that I encounter a white marble statue that is bathed in green light in such a way as to make the statue appear green—a workaday perceptual illusion. Suppose further that I do not believe that the statue really is green. I have seen the statue under normal lighting conditions, I know that no one has tampered with the statue (for instance, by painting it green), and I can see the source of the green light. Despite all this, could the statue's (merely) appearing green prompt me to change my mind about what color the statue really is?

I do not see how it could. Absent other factors—a 'credulousness' drug, reasonable doubts about whether the statue was modified before I entered the room, and so forth—I do not see how the perceptual illusion could prompt me to revise my belief. For that matter, I do not see how the perceptual illusion could even *tempt* me to revise my belief.

The point is not that perceptual illusions always fail to make us revise our beliefs. They succeed in this quite often, of course. But never when we recognize them as illusions. Even illusions that are vivid and terrifying cannot make us revise our beliefs if we know that they are illusions. I once went on a ride at Disney Land that used holographs to give riders the sensation that they were hurdling toward a head-on col-

lision with a speeding train. Although I flinched reflexively, I did not believe one bit that the train was really there. Few if any of the cognitive hallmarks of belief were present: I was not disposed to assert that the train was really there, for example, nor did I treat the proposition that the train was really there as a premise in further reasoning. The holographic image of the train only *seemed* to be a reason to believe that the train was hurdling toward me, and so I did not form the belief that it really was.

These reflections suggest that the standard evaluative account of akratic action, with its analogy between akratic action and illusion, is not just unhelpful to the evaluative conception; it is fatal. Far from showing how akratic action is intelligible, the standard evaluative account appears to entail that akratic action is impossible. For if a known perceptual illusion cannot tempt or cause belief, then—taking seriously the analogy between evaluative appearance and perceptual illusion—it would seem to follow that a known illusory evaluative appearance cannot tempt or cause action. In particular, it would seem to follow that, no matter how good or right or justified some prospective act of violent retaliation might appear, such an illusory appearance could not tempt me to retaliate. The basic point is this: a thing's seeming good (when I know it is not) can no more tempt me to perform an action than a thing's seeming green (when I know it is not) can tempt me to form a belief.

When I assert that a thing's merely seeming to be a reason cannot by itself motivate me to revise my prior beliefs, I do not mean to deny the possibility of epistemic *akrasia*. (Epistemic or theoretical *akrasia* is the cognitive condition in which our beliefs are somehow out of line with our judgments about what we ought to believe—say, because we form beliefs that we think the evidence doesn't support or because we fail to form beliefs that we think the evidence does support.) As T.M. Scanlon imagines,

> I may know that despite Jones's pretensions to be a loyal friend, he is in fact merely an artful deceiver. Yet when I am with him I may find the appearance of warmth and friendship so affecting that I find myself thinking, although I know better, that he can be relied on after all.[32]

[32] Scanlon, What We Owe, supra note 7, at 35.

This case seems possible but under-described. For I cannot make sense of my revising my belief about Jones if we suppose that no factor other than Jones's "appearance of warmth and friendship" contributes to my revising the belief. If I know Jones's apparent warmth and friendship to be a sham (in effect, an illusion) it seems to follow that Jones's apparent warmth and friendship cannot (other things absent) prompt me to *believe* him to be trustworthy—just as the statue's seeming green cannot (other things absent) prompt me to *believe* the statue to be green. Given my prior conviction to the contrary and the manifest worthlessness of the 'evidence' with which I am now presented, there is no plausible psychological explanation for my change in belief.

In order to explain the change, we must posit an additional element—perhaps a desire that Jones be a genuinely trustworthy friend, or even just a desire for friendship.[33] Either desire could make good sense of my changing my opinion about Jones, as wishful thinking is common. But we cannot understand such a desire as an evaluative appearance, for my being in the grip of a desire *qua* evaluative appearance is no more than a thing's seeming to present me with a reason, and a thing's merely seeming to present me with a reason cannot (as we just saw) motivate me all by itself to revise my prior beliefs—especially not when the prior beliefs are based on what I take to be credible evidence. What makes possible this particular variety[34] of epistemic *akrasia*, I suggest, is something more than the fact that I encounter misleading evidence; it is the fact that the misleading evidence I encounter tempts me to believe what I antecedently *wanted* to believe. A necessary ingredient of wishful thinking, after all, is that there is something that I *wish* for, something that I *want*. But then it is not my evidence that tempts me; it is my desire.

[33] I don't mean to suggest that desires are the only possible causes of epistemic *akrasia*. They are probably not even the most common causes. For helpful discussions of epistemic *akrasia*, see Tenenbaum, The Judgment of a Weak Will, supra note 7; David Owens, Epistemic Akrasia, 85 Monist 381 (2002).

[34] Wishful thinking is of course but a single variety of epistemic *akrasia*. Other varieties are made possible by other mechanisms. What's more, wishful thinking need not be akratic, because the desire that motivates an instance of wishful thinking might motivate me to revise not only my first-order belief but also my second-order belief about what first-order belief is warranted by the evidence, thereby 'curing' my *akrasia*.

"But perceptual illusion really *can* tempt belief," an evaluative theorist might insist. "When I see a straight stick that is partially submerged in a bucket of water, I am ever so tempted to believe the stick is bent, even though I know very well it is not."

Then I will grant for the sake of argument that perceptual illusion can *tempt* belief. My granting this, however, will do nothing to establish that any perceptual illusion, no matter how tempting, can actually *cause* belief. However tempted you may be, you surely will not come to believe that the stick is really bent. Nothing could be easier than resisting this *doxastic* (belief-related) temptation. And if nothing could be easier than resisting this doxastic temptation, then—taking seriously the analogy between evaluative appearance and perceptual illusion—it would seem to follow that, contrary to the first truism, nothing could be easier than resisting a *practical* temptation. But if that were true, we would find it entirely mysterious—indeed unintelligible—how anyone could ever act akratically.[35]

"Whether or not perceptual illusion can exert an influence on belief," an evaluative theorist might reply, "evaluative appearance *can* exert an influence on intention and action. You have not proved otherwise."

But it is not true that evaluative appearance can exert an influence on action, and I can show this without exploiting the analogy that we have now supposed the evaluative theorist to abandon—the analogy between evaluative appearance and action, on the one hand, and perceptual illusion and belief, on the other. Imagine that as I am preparing to shave one morning, I squirt a large dollop of shaving cream into my palm. I am suddenly struck by the fact that this dollop of shaving cream looks exactly like a luscious dollop of whipped cream. Am I tempted to eat it? No. If I were tempted, would I succumb? Of course I would not.

"But that's because you know it would taste awful and make you sick."

Fine. Instead of a dollop of shaving cream, it is a visually indistinguishable dollop of imitation shaving cream, a flavorless concoction of egg whites that would not taste awful or make me sick. Am I

[35] This might help explain why evaluative theorists from Aristotle to Donald Davidson have found akratic action so puzzling.

tempted to eat it? If I were tempted, would I succumb? As delicious as such a dollop might *look*, I can hardly imagine that I would be tempted to eat it, let alone that I would actually do so.

"But these examples miss the point. They prove nothing because they are examples of *perceptual* illusions (loosely speaking), not of *evaluative* illusions. To be sure, the dollop of imitation shaving cream does appear as though it would taste a certain way (i.e., like whipped cream). But no evaluative theorist ever claimed that temptation leads us astray by presenting a thing as having a non-evaluative property (e.g., a certain flavor) that the thing actually lacks. The claim is that temptation leads us astray by presenting a thing as having an *evaluative* property that the thing either lacks outright or possesses to a lesser degree. Forget about imitation shaving cream. Think about real whipped cream. Real whipped cream appears delicious. Now this appearance is not an illusion; whipped cream really is delicious. The illusion is that the deliciousness of the whipped cream appears *better* than it really is; it appears to give us a stronger practical reason (to eat the whipped cream) than it really gives us. The underlying practical reason is real, of course, because (other things equal) a thing's being delicious really is a reason to eat it. But this reason appears stronger than it actually is—stronger, in fact, than we know it to be. And it is *that* illusory appearance—not any other—that leads us astray."

The evaluative theorist's position is now a good deal clearer, but it is also even less plausible. The evaluative theorist now posits that evaluative illusion has a kind of motivational efficacy that non-evaluative illusion lacks. But how could that be? How could I be motivated by the fact that the whipped cream's deliciousness seems good, yet not be motivated by the fact that the imitation whipped cream seems delicious? Again, the analogy to perceptual illusion is not just unhelpful; it is damaging. If there is anything to be learned from the analogy, it is this: from the fact that non-evaluative illusion is motivationally inefficacious, we might very well infer that *evaluative* illusion is motivationally inefficacious, too.

In response, an evaluative theorist might insist that evaluative illusion is distinctive in just this way, that it has a kind of motivational efficacy that non-evaluative illusion lacks. But if evaluative illusion really has this kind of motivational efficacy, we should ask why evaluative illusion nevertheless lacks *doxastic* efficacy, why it can make me

*do* something without at the same time making me *believe* something. For it would seem, on the contrary, that any evaluative appearance capable of motivating me to act contrary to my all-things-considered judgment should also be capable of getting me to revise that judgment, if the mechanism of temptation is nothing but practical illusion. (How could *X*'s seeming good cause me to pursue *X* without at the same time causing me to believe *X* to be better than *Y*? How could an apparent reason be so persuasive as to get me to move my limbs yet be unable to get me to change my mind?) But if any temptation capable of motivating me to act against my all-things-considered judgment were also capable of getting me to revise that judgment, temptation would never give rise to action that was truly akratic. Temptation could lead me astray only by deceiving me. My succumbing to temptation might be irrational, then, but it would never be akratic, because in succumbing to temptation I would necessarily be changing my evaluative beliefs to fit my actions. Any evaluative theorist who disagrees must shoulder the burden of explaining what mechanism could persuade me to act as though the world were a certain way without thereby persuading me to represent the world as being that way. Perhaps this burden can be discharged, but I do not see how. I find it hard to understand how an evaluative appearance could motivate someone to perform an action without at the same time motivating him to revise his evaluative beliefs.

Because it renders *akrasia* impossible or at least unintelligible, the evaluative conception evidently cannot undergird the defense of provocation, if the prototypical act of retaliation is both akratic and intelligible. (And what could be more intelligible than an act of vengeful retaliation? What could *make more sense* than this—both from the standpoint of the actor and from the perspective of an onlooker?) The problem is actually deeper, of course, because the reason why the evaluative conception renders akratic action unintelligible is that it defies the first truism about temptation, making it seem as though resisting temptation should be as easy as refraining from endorsing the content of an obvious illusion. Thus, even if the prototypical case of retaliation is not in fact a case of akratic action, the evaluative conception is still inconsistent with all of the major volition-based defenses. Each of these defenses—provocation, duress, intoxication, and insan-

ity—contemplates a defendant who finds resisting some criminally-directed temptation to be highly difficult, if not practically impossible.

The evaluative conception fails as a theory of the criminal law's volition-based defenses for what is ultimately a simple reason: it cannot accommodate an unassailable truism, that resisting temptation is ordinarily difficult. Unable to accommodate this truism, the evaluative conception fails also as a theory of temptation.[36]

## III. The Mechanistic Conception of Temptation

The evaluative conception founders on its inability to explain volitional impairment. The mechanistic conception promises to do better because it identifies temptation with something quite similar to volition itself. As we will see, the mechanistic conception does manage to explain volitional impairment. But the mechanistic conception makes temptation *difficult* to resist at the cost of making it *impossible* to resist, saving the first truism about temptation by eviscerating the second. The second truism, I have suggested, helps explain why provocation and duress are *incomplete* defenses, in that provocation mitigates without fully exculpating, and duress fully exculpates but does not apply to all crimes.

The mechanistic conception of temptation shows up in the work of Descartes, Kant, and Freud, as well as in the behaviorist and neo-behaviorist psychology of the twentieth century.[37] Its most ardent con-

[36] Untouched by my argument are those extreme varieties of the evaluative conception that equate temptation not with evaluative appearance but with full-fledged evaluative judgment. (For theories of emotion along these lines, see, e.g., Solomon, supra note 7; Neu, supra note 7; Nussbaum, supra note 7.) These theories share many of the virtues of their less cognitive cousins, offering insightful accounts of the nature and moral significance of the emotions, but are vulnerable to the objection that temptation *qua* evaluative judgment is incapable of underwriting genuine akratic action. If a person acts on a temptation *qua* evaluative judgment that conflicts with his 'dispassionate' practical judgment—that is, his *belief* about what he ought to do—then he lacks a wholehearted all-things-considered judgment. But if he lacks a wholehearted all-things-considered judgment, he is not akratic so much as ambivalent.

[37] See, e.g., René Descartes, Passions of the Soul; Sigmund Freud, A General Introduction to Psycho-Analysis (1935). See Brewer, supra note 29, for a discussion of Kant's mechanistic account of desires. See Kahan and Nussbaum, Two Conceptions of Emotion, supra note 7, at 280-82, for a concise history of the mechanistic conception of emotion, replete with references to contemporaries and historical figures.

temporary defender is probably the philosopher Harry Frankfurt, who explains the mechanistic conception like this:

> However imposing or intense the motivational *power* that the passions mobilize may be, the passions have no inherent motivational *authority*. Considered strictly in themselves, apart from whatever additional impetus or facilitation we ourselves may provide by acceding to them, their effectiveness in moving us is entirely a matter of *sheer brute force*. There is nothing in them other than the magnitude of this force that requires us, or that even encourages us, to act as they command.[38]

Elsewhere Frankfurt advances a mechanistic conception specifically of "our most elementary desires":

> Animals of many species have desires, but only animals of our species—or, perhaps, of a few others—are capable of seeing anything as a reason. Our most elementary desires come to us as urges or impulses; we are moved by them, but they do not as such affect our thinking at all. They are merely psychic raw material. A desire provides us not with a reason but with a problem—the problem of how to respond to it. Impulses and urges have power, but in themselves they have no authority. They move us more or less strongly, but they make no claims on us.[39]

So conceived, temptations consist solely in *felt motivation*: "They are merely psychic raw material" whose "effectiveness in moving us is entirely a matter of sheer brute force." For this reason, temptations mechanistically conceived have no normative authority: they do not in themselves give rise to practical reasons or requirements. They in fact lack even the mere pretense of normative authority: they do not even seem to give rise to practical reasons or requirements. Nor do they present their objects as good or valuable. Indeed, not only do temptations mechanistically conceived fail to present their objects as good or valuable, they fail to present their objects even as appealing or attractive: "There is nothing in them other than the magnitude of [their sheer brute] force... that even encourages us... to act as they command."

[38] Harry G. Frankfurt, Autonomy, Necessity, and Love, *in* Frankfurt, Necessity, Volition, and Love (1999), at 137.

[39] Harry G. Frankfurt, Reply to T. M. Scanlon, *in* Buss and Overton, supra note 7, at 184.

Temptation *qua* vector of sheer brute force being the natural enemy of volition, it is not difficult to see how temptation mechanistically conceived might be thought to undergird the volition-based criminal defenses. But how might temptations conceived as vectors of sheer brute force lead us to act akratically?

One possibility—which in the end I will argue is the only true possibility—is that these vectors of sheer brute force lead us to act akratically by being overpowering compulsions, impulses that are literally irresistible. This possibility is of course a non-starter for anyone who would care to remain faithful to the second truism about temptation, that no matter how difficult resisting temptation sometimes might be, resisting is not ordinarily impossible. So we should take a moment to reflect on why this truism is worth accepting. Some accept the truism because they assume that, whenever we act on the basis of an irresistible impulse, we exhibit some kind of pathology. This assumption seems false, however. Under no circumstance could I refrain from diving in a lake to save my drowning child. The desire to do so is irresistible if any desire is. But that hardly makes my action pathological.[40] So the problem with supposing all temptations to be overpowering compulsions is not that when we act from an irresistible impulse we exhibit some kind of pathology. The problem is simply that most temptations are not in fact irresistible. The temptation to retaliate against an outrageous provocation, for one, is not ordinarily irresistible. Even genuinely pathological compulsions are not ordinarily irresistible. Pyromaniacs reliably refrain from compulsive behavior when the threat of arrest is imminent. Alcoholics abstain for weeks when their circumstances demand it. People with obsessive-compulsive disorder readily distract themselves from persistent compulsions using learned visualization techniques. But if even pathological urges are generally resistible, surely so are ordinary temptations.

So maybe temptation *qua* vector of sheer brute force leads us astray not through compulsion but through what we might call "diversion." Just as an extra-personal force like a gust of wind can be strong

[40] Compare the discussions of "volitional necessity" in Harry G. Frankfurt, The Importance of What We Care About, *in* Frankfurt, The Importance of What We Care About (1982), *and* Gary Watson, Volitional Necessities, *in* Watson, Agency and Answerability (2002).

enough to dissuade me from forging ahead without being so strong as to blow me off my feet, so might an intra-personal force be strong enough to divert me from my chosen course without wholly compelling me to give it up. If this is the mechanism by which temptation operates, then I act on a temptation (when I do) not because I can't resist it—I can—but because I *don't* resist it. And why don't I? Because I choose not to; because I acquiesce. So when I strike back against my provoker, I do so not because my retaliatory desire overpowers me, but because I yield to its power.

This 'diversion' model is an improvement on the 'compulsion' model, in so far as the former does not depict temptations as overpowering compulsions. Yet the 'diversion' model cannot but render akratic action less than full-blooded—not because it renders akratic action involuntary but because it renders such action mysterious, indeed, unintelligible. If a temptation *qua* vector of sheer brute force is resistible, then why do I not resist it? You will strain to make sense of my acquiescence and, crucially, so will I. Conceived as a vector of sheer brute force, a temptation does not entice or seduce, does not present its object as appealing or attractive. Much less does it present its object as good or valuable. That I would act on such a temptation in defiance of my all-things-considered judgment is bizarre. It is something I will not be able to make any sense of—something I will find unintelligible. The problem here is not the puzzle of perversity, the puzzle of how a person could intelligibly pursue the bad for the sake of nothing but its badness.[41] That is the puzzle of how a person could intelligibly find the bad appealing. The present problem is worse. It is the problem of how a person not acting on an irresistible compulsion could, in defiance of his all-things-considered judgment, intelligibly pursue what he finds in no way appealing.

[41] Some philosophers have found the puzzle of perversity to be insoluble and accordingly have embraced the view that no one intentionally pursues the bad. See, e.g., Aristotle, Nicomachean Ethics 1094a1-3, *in* The Complete Works of Aristotle, supra note 7; St. Thomas Aquinas, The Summa Theologica (The Fathers of the English Dominican Province trans., 1947), QQ. 8, available at http://www.ccel.org/a/aquinas/summa/home.html; Anscombe, Intention, supra note 9, at 73-74; Dennis W. Stampe, The Authority of Desire, supra note 7, at 355-81; Tenenbaum, Appearances of the Good, supra note 7.

I should emphasize that I am not simply rehearsing the well-worn charge that "a mere behavioural disposition [cannot] make the behaviour it causes intelligible,"[42] a charge that philosophers customarily have substantiated by appealing to examples like that of Warren Quinn's 'radio man,' a fantastical character beset by an inexplicable, affectless urge to turn on every radio he encounters.[43] For my part, I do not share the intuition that such affectless urges are by their nature apt to cause behavior that is unintelligible. On the contrary, I find it rather obvious that people often act intelligibly on brute urges, urges that do not depict their objects as appealing or attractive. People hum tunelessly under their breath, drum their fingers on the table, avoid stepping on cracks in the sidewalk—all because of brute urges. None of this is unintelligible or even out of the ordinary. But what makes this behavior intelligible and ordinary, I suggest, is that it is harmless. None of it contravenes the actor's all-things-considered practical judgment. If, on the other hand, a person thinks it harmful to waste his energy on avoiding the cracks in the sidewalk yet perseveres in this avoidance solely because of a resistible brute urge, his behavior surely will be unintelligible—not just to others but to himself as well. For it is utterly mysterious why a person not acting on an irresistible compulsion would, in defiance of his all-things-considered judgment, pursue what he does not find in any way appealing. The only thing that could make this behavior intelligible, it seems, would be for the temptation to be irresistible; only if the impulse were irresistible would it make any sense for the person to acquiesce. But if a person could act contrary to his better judgment only when afflicted by an irresistible impulse, then he hardly could act contrary to his better judgment *voluntarily*.

A mechanistic theorist might object to all this as follows:

> You have assumed that resisting a temptation is always costless. But your assumption is false. Far from being costless, resisting a temptation may be difficult or taxing or painful. We may choose not to resist,

[42] Tappolet, Emotions and the Intelligibility of Akratic Action, supra note 29, at 112. On this point, Tappolet cites: Warren Quinn, Putting rationality in its place, *in* Quinn, Morality and Action (1993), at 236-37; Scanlon, What We Owe, supra note 7, ch. 1; Jonathan Dancy, Practical Reality (2000), ch. 2; Johnston, supra note 7.

[43] Quinn, supra note 42, at 236-37.

> then, simply because we don't want to go through the trouble. What makes such a choice akratic is that, by our own lights, the difficulty of resistance is not great enough to justify our acquiescence. We acquiesce not because we deem resistance unjustifiably difficult—we don't—but because we are lazy or fearful or selfish. This choice may be unwise or immoral (although it need not be). But it surely isn't unintelligible.

That is a reasonable point, but it is a point that no mechanistic theorist can make, for the mechanistic conception applies no less to a person's aversion to the discomfort of resisting a given temptation than it applies to the underlying temptation itself. The aversion is but another temptation, so the aversion, too, must be a (resistible) brute urge. But if the aversion is itself a resistible brute urge, then the problem of unintelligibility will simply reassert itself. If we cannot intelligibly acquiesce to the brute urge that is the underlying temptation, neither can we intelligibly acquiesce to the brute urge that is our aversion to resisting the underlying temptation. Only if the aversion were irresistible could we acquiesce to it intelligibly.

In effect, the mechanistic conception supposes that temptation cannot cause us to act akratically unless it overpowers us. This means that the mechanistic conception accommodates the first truism at the cost of junking the second; it makes temptation difficult to resist by making it impossible to resist. Accordingly, the mechanistic conception conflicts with (at least) the defenses of provocation and duress, which presume that the defendant's temptation is not altogether irresistible.

Faithful to the mechanistic conception in spirit, someone might try to circumvent these problems by adopting a hybrid view, specifically, a mechanistic conception of (most) temptations coupled with a non-mechanistic conception of the aversion to resisting these temptations. Such a hybrid view would have the obvious virtue of being able to make sense of (that is, render intelligible) two common varieties of akratic action: (i) that motivated by a temptation that it would be difficult or unpleasant to resist and (ii) that motivated by a temptation the mere possession of which is unpleasant or painful. Applying the hybrid view to the case of provocation, we would regard the provoked defendant either (i) as someone who finds his retaliatory desire unusually difficult or unpleasant to *resist* or (ii) as someone who finds his

retaliatory desire especially unpleasant or painful to *possess*, and who therefore acts on that desire in order to extinguish it.

Despite its ability to accommodate limited varieties of akratic action, the hybrid view is bound to fail as an overall strategy because most cases of akratic action—including the action of retaliating lethally against a provoker—have a quite different character. As Gary Watson has observed, temptation typically functions less like a bully than like a seducer.[44] This is a fact the hybrid view cannot accommodate. The hybrid view supposes that temptation always assails us with a threat: if we don't satisfy it, it will make us pay. Temptation functions this way sometimes, perhaps, but not always. Not even close to it. Much more often, temptation approaches us not with a threat but with a promise, the promise of some forbidden indulgence.

This difference—between threat and promise, bully and seducer—is not simply phenomenological. Yielding to a psychic seducer differs from yielding to a psychic threat in terms of fundamental motivational structure. When we yield to a threat (say, the urgent desire to scratch an itch) our motivation has a dual basis. What moves us is not just the threat but also our aversion to the anticipated discomfort of resisting it: we are drawn to the (future) pleasure of scratching just as much as we are repulsed by the (present) discomfort of *not* scratching. But when we yield to a seducer (a typical gustatory desire, for example) our motivation has no such dual basis. What moves us is just the seducer. It would be redundant to posit a further desire for the anticipated pleasure of acquiescence, a desire analogous to our aversion to the anticipated discomfort of resisting a psychic threat. A desire for the anticipated pleasure of acquiescing to a psychic seducer would be identical in every respect to the psychic seducer.

Which of these structures better fits provocation—psychic threat or psychic seducer? If the retaliatory desire is a psychic threat, the provoked actor is motivated not just by the retaliatory desire but also by his aversion to resisting the desire. This means that the actor's attention is directed both at the retaliatory desire's object—the pleasure of violent retaliation—and at the desire itself: not only does he attend to the appealing prospect of violent retaliation, but he also attends to

[44] Gary Watson, Disordered Appetites: Addiction, Compulsion, and Dependence, *in* Watson, Agency and Answerability, supra note 40, at 63-66.

the imminent pain and difficulty of resisting his desire for that appealing prospect. If the retaliatory desire is instead a psychic seducer, the provoked actor is motivated just by that desire, his attention being directed exclusively to the desire's object: the appealing prospect of violent retaliation. The second structure strikes me as more accurate to the phenomenon of provocation. The provoked actor is not someone preoccupied by his own psychic economy, the way a recovering alcoholic might be preoccupied by his desire for drink. The provoked actor's attention is focused very much outward, on the enticing prospect of retaliation.

This difference in motivational structure underwrites the corresponding difference in phenomenology. Resisting a psychic threat always feels hard—very hard, if the threat is onerous or if we are unusually lazy or fearful or selfish. Not so resisting seduction. Though it may be difficult to resist the seduction of revenge, seductive temptations are in some cases quite easy to resist, and shamefully so. Imagine a person who finds himself on the verge of committing a sexual indiscretion that he could easily avoid. With minimal difficulty, he could extricate himself from his tempting predicament and take a brisk walk around the block. Taking the walk would be mildly pleasant in itself (since it is a beautiful summer evening, let's suppose) and the distraction would extinguish his sexual desire. Now imagine that, despite the ease with which he could resist temptation, the person opts for sex. Must we say that he opts for sex because he is unwilling to go through the trouble of resisting his sexual desire? I don't see why we must. For we have stipulated that this particular desire is easy (even somewhat pleasant) to resist and eradicate. The person opts for sex, then, not because resistance is difficult or unpleasant but because acquiescence is appealing. Certainly, some temptations are of such a character that acting on them feels like giving in to a threat. But most don't have this character. Most approach us not as threats but as seducers. We succumb not because resistance would be futile but because it would be no fun.

## IV. A Conjunctive Theory of Temptation?

If we would be true to the criminal law and the conception of human action that it presupposes, we must reject any theory that identifies the

essence of temptation either with felt motivation or with evaluative appearance. For all I have said, however, there is nothing else we must reject. Motivation and evaluation still might be central to temptation; indeed, they might be its essence jointly.

On a 'conjuctive' account of temptation, the desire for *X* will consist in the coupling of (i) a raw motivation to pursue *X* with (ii) the (quasi-perceptual) appearance of *X* as good or valuable. Can a conjunctive account avoid the defects that plague each of the conjuncts? Can motivation and evaluation solve each other's problems? You will probably think they can if you think that

(1) a brute urge can make an evaluative appearance motivationally efficacious without rendering a person's subsequent behavior unintelligible.

On the other hand, you will probably deny that motivation and evaluation can solve each other's problems if you think that

(2) a brute urge can make an evaluative appearance motivationally efficacious only at the cost of rendering a person's subsequent behavior unintelligible, such that he will ask himself, "Why am I behaving as though this obvious illusion were real? Why am I chasing this mirage? What is this force that impels me?"

I am inclined to agree with (2). I cannot see how your acting on an evaluative appearance, even at the prompting of a brute urge, could be any more intelligible than your believing the straight stick to be bent. If you experienced a 'brute doxastic urge' to believe that the stick was bent and on that basis came to believe that it really was, you surely would ask yourself, "Why am I endorsing the content of this obvious illusion? What is this force that impels me?" Similarly, if you acted on a brute urge in the service of what you knew to be an illusory evaluative appearance, I submit that you inevitably would ask yourself, "Why am I chasing this mirage? What is this force that impels me?"

Now, you might agree that these last two questions are inevitable but go on to insist that their inevitability is desirable, not damning, for you might think that these questions capture something that any account of temptation must capture, namely, the fact that akratic

action characteristically involves feelings of alienation and even of borderline unintelligibility—feelings very nicely expressed by the questions, "Why am I chasing this mirage? What is this force that impels me?"

I would be the last to deny that akratic action often involves these feelings. My point is that akratic action involves other feelings, too—feelings of intelligible attraction and of affect-laden motivation. These feelings explain why the non-optimal path is often the seductive and alluring one, why akratic action is often at least partly intelligible.

What, then, of these feelings of intelligible attraction and affect-laden motivation? Do these feelings admit of a unitary theory, along the lines of the mechanistic and evaluative conceptions of temptation? They may. But I doubt it. It does not seem likely that states as diverse as anger, fear, lust, hunger, hatred, and fatigue have in common some single feature (or set of features) in virtue of which each is apt to lead us astray. Now, these states do have *some* features in common, of course, and that explains a principal attraction of the mechanistic and evaluative conceptions, namely, their success in identifying several of these common features. For its part, the mechanistic conception is surely correct that all temptations bear an essential relation to felt motivation. (Hunger involves a motivation to consume food, and fatigue involves a motivation to rest as well as the conspicuous absence of a motivation to engage in strenuous activity.) The evaluative conception is equally correct that many paradigmatic temptations involve evaluative appearances, if not full-fledged evaluative beliefs. (Fear involves a sense that one is in danger and should flee, and anger sometimes involves a sense that one has been wronged and should retaliate. As Aristotle says, we learn "that we have been insulted or slighted, and anger, reasoning as it were that anything like this must be fought against, boils up straightway."[45]) Where the mechanistic and evaluative conceptions go wrong, then, is not in supposing that there is some feature common to paradigmatic temptations. Where they go wrong is in supposing that this common feature, be it evaluation or felt motivation, is the essence of temptation, the thing in virtue of which a temptation leads us astray. Evaluation and felt motivation may well be

[45] Aristotle, Nicomachean Ethics 1149a31-33, *in* The Complete Works of Aristotle, supra note 7.

central to paradigmatic temptations, but neither can be what leads us to act akratically. Neither can be what makes temptation tempt.

None of this means that we must give up trying to make sense of temptation. We simply should scale back our theoretical ambitions and proceed piece by piece, taking one variety of temptation at a time. That is precisely the strategy I will adopt in the remainder of this essay. First, I will sketch an alternative theory of one important variety of temptation: affective desire. Then I will show how this theory can make sense of the retaliatory desire presumed by the provocation defense.

## V. A Theory of Affective Desire

There is something peculiar, almost paradoxical, about what it's like to resist temptation: it is at once all too easy and all too hard. In hindsight, it is all too easy; hence our frustration and our self-reproach. In the moment, it is all too hard; hence our chronic failure to stand firm. That temptation has this peculiar phenomenology is something a theory of desire should accommodate, if not explain. The prevailing theories do neither. What we need is a theory of desire that accommodates not just the possibility of succumbing to temptation but also temptation's particular phenomenology. We need a theory that does justice to how desire leads us astray.

Desire leads us astray, I will argue, by presenting us with a kind of primitive (that is, non-analyzable, indefinable) felt attraction. This attraction differs from motivation and also from evaluation, two psychological phenomena with which desire is almost always coupled in practice but with which it should never be confused in principle. If we confuse attraction with motivation, we will depict temptation as a contrary causal force, a force to which we succumb (when we do) by surrendering in a psychic tug-of-war. As we saw in Part III, this conception of temptation makes resistance out to be harder than it really is. Although succumbing to temptation sometimes has the character of surrendering at tug-of-war, more often it is a matter of indulging in a forbidden pleasure. This feels different from acquiescing to a contrary causal force. It feels less like compulsion than like seduction. If we confuse attraction instead with evaluation, we will depict temptation not as a form of seduction but as a mode of quasi-rational persuasion.

As we saw in Part II, this conception of temptation makes resistance out to be easier than it really is. If akratic temptations are but evaluative illusions—illusions, moreover, that we know to be just that—resistance should be easy. Resistance should be so easy, in fact, as to be ubiquitous. Yet it obviously isn't ubiquitous. And it often isn't easy. What we need, then, is a theory of desire that (unlike the mechanistic conception) allows for *affect*, but (unlike the evaluative conception) does not reduce affect to *appraisal*. I offer such a theory here.

Philosophers at one time or another have attached the label 'desire' to just about every kind of motivating state. My target is considerably narrower. As Mark Johnston observes,

> there is a perfectly good non-philosophical sense of 'desire' in which desire is not only one of the springs of action, but a state which makes certain kinds of actions readily intelligible. In this sense of 'desire', which we might distinguish by the somewhat pleonastic name 'affective desire', we desire other things and other people, we are struck by their appeal, we are taken with them.[46]

Desire in this sense is a thing we fight, indulge, and condemn; a thing that makes our lives colorful, meaningful, and difficult. Though there are other kinds of motivating state that philosophers have sometimes called by the name 'desire'—those arising from our decisions, intentions, and practical judgments, for example[47]—these other motivating states fall outside the scope of my inquiry. My inquiry pertains more narrowly to affective desire. A prime exemplar of affective desire is of course the urge to retaliate against an outrageous provocation.

My chief contention is that the essence of affective desire (hereafter, simply 'desire') is a feeling of *psychic attraction*. As I will argue, this feeling is not reducible to any combination of the following: a feeling of motivation; an appearance of the desired object as good or valuable; an ascription of appealing properties to the desired object; a state of insistent attention to the desired object's appealing properties. These phenomena ordinarily accompany desire, but none is the thing itself.

[46] Johnston, supra note 7, at 188. Although I agree with much of what Johnston says about the character of affective desire, I reject his core claim that such desire consists in the perception of a certain kind of value.

[47] See, e.g., Thomas Nagel's discussion of "motivated desires" in The Possibility of Altruism 29 (1970).

I suggest that the feeling of psychic attraction is itself indefinable, that it can't be analyzed in terms of more primitive psychic phenomena. Here I take a page from David Hume, who made a similar observation while discussing pride and humility:

> The passions of PRIDE and HUMILITY being simple and uniform impressions, 'tis impossible we can ever, by a multitude of words, give a just definition of them, or indeed of any of the passions. The utmost we can pretend to is a description of them, by an enumeration of such circumstances, as attend them...[48]

Something similar seems true of desire: though we may not be able to define it, we can describe the phenomena that ordinarily attend or characterize it. In this respect, desire is no different from a host of familiar psychological states. The emotions each have a distinctive but ineffable feel, and so do the many varieties of pain and pleasure. Though there is much we can say about the cognitive and volitional conditions that characteristically attend these psychological states, there is little we can say about the states' felt quality—little that is not hopelessly metaphorical. So instead of *describing* psychic attraction, I will *isolate* it—distinguishing it from the phenomena by which it is typically accompanied and for which it is often confused.

We can make some headway in isolating psychic attraction if we start by scrutinizing a single case. Consider thirst. In adult humans, thirst is a syndrome that involves at least an experiential state (an unpleasant sensation of dryness in one's mouth and throat) and a thought (the thought that a certain activity would relieve the unpleasant sensation).[49] Crucially, thirst also involves a feeling of attraction—whether to drinking (an act), to potable liquid (a thing), or to one's thirst being slaked (a state of affairs). For our purposes, it doesn't matter how we conceive of thirst's object. What matters is how we understand the relevant feeling of attraction.

A natural thought is that we should understand the feeling of attraction as a kind of felt motivation. This thought derives some initial plausibility from the obvious fact that a thirsty person's desire for

[48] David Hume, A Treatise of Human Nature (L.A. Selby-Bigge and P.H. Nidditch eds., 1978), at 277.

[49] Scanlon, What We Owe, supra note 7.

drink grounds a set of motivational dispositions, for example, dispositions to seek out potable liquids and to consume them. But it is a further question whether a thirsty person's felt attraction to drinking is nothing but the conscious manifestation of these motivational dispositions. The answer, I suggest, is no. Felt motivation seems insufficient for psychic attraction, as I might feel myself motivated to pursue things to which I am not the least bit attracted. Suppose that, suffering from a disease like pica, I am afflicted by a brute urge to consume certain vile-tasting fluids, such as glue and liquid laundry detergent.[50] Though I find these fluids disgusting, I nevertheless am beset by a resistible but insistent impulse to drink them. My urge hardly would seem to be a case of desire, as there is no respect in which I find the prospect of drinking *appealing*.

This is by no means to suggest that brute urges are always bizarre or pathological.[51] I may experience a perfectly healthy brute urge to fidget with my wedding ring or to click my pen repeatedly. When I do these things, I do them simply because I feel motivated to do them, not because the actions strike me as appealing. Now the impulses that prompt these actions surely are not pathological. They are ubiquitous, innocuous, and normal. But that doesn't make these urges desires. When I desire something, I am not merely motivated to pursue it. I am *taken* with it. I find it appealing.

What is it, then, to find something appealing, to be psychically attracted to it? If finding something appealing is not a matter of feeling a motivation to pursue the thing, is it instead a matter of the thing seeming good or valuable? When we desire something, it almost always strikes us as good in some respect and to some degree, even if we simultaneously judge that the thing is not actually good at all. Indeed, it might well be true that we can desire something *intelligibly* only if we find it good in some way or other.[52] It is certainly natural to say of the things we desire that they "seem good."[53] But if being psychically

[50] For a description of pica, see American Psychiatric Association, The Diagnostic and Statistical Manual of Mental Disorders (4th ed.) Text Revision 103-105 (2000).

[51] Here I disagree with Johnston, supra note 7, at 190, who asserts that "affect is close to ubiquitous."

[52] Anscombe, Intention, supra note 9; Raz, Incommensurability, supra note 9; Raz, Agency, supra note 9.

[53] Hawkins, supra note 7, at 244.

attracted to something were simply a matter of finding it good, psychic attraction could not underwrite akratic action (for all the reasons we considered in Part II).

In any event, further consideration of the phenomenology of desire yields two more reasons to doubt that psychic attraction is just evaluative 'seeming.' The first is that what appeals to us is not always the same as what strikes us as good.[54] Finding a thing good is one way in which you can find a thing appealing. But it is not the only way. As Augustine observed long ago, an action can appeal to you in certain cases precisely because it strikes you as *not* good.[55] This much is implicit in the commonsense understanding of such phenomena as spite. To be spiteful is to be attracted to the prospect of harming someone even when doing so seems wholly bad—indeed, precisely *because* doing so seems wholly bad.

But you may doubt the possibility of our being attracted to the bad as such[56] and so you should consider a second and perhaps deeper reason not to equate finding-a-thing-appealing with a-thing-seeming-good: even when something simultaneously appeals to us *and* strikes us as good, our finding the thing appealing seems on inspection to be psychologically distinct from the thing's striking us as good.[57] When I am thirsty and desire water, I mentally represent the (prospective) action of drinking in certain characteristic ways, imagining for example the sensation of a cool liquid passing through my lips and down my throat. Representing in this fashion the experience of drinking is ordinarily a way of taking the experience to be pleasant. But it is not in itself a way of taking the experience to be good. No doubt the pleasure of drinking is in fact a good thing. This, however, is a *further* fact, a

[54] Saint Augustine, The Confessions (Philip Burton trans. and ed., 2001); Michael Stocker, Desiring the Bad: An Essay in Moral Psychology, 76 J. Phil. 738 (1979); Michael Stocker, Plural and Conflicting Values (1990); Michael Stocker, Raz on the Intelligibility of Bad Acts, *in* Reason and Value: Themes from the Moral Philosophy of Joseph Raz (R. Jay Wallace, et al, eds., 2004); David Velleman, The Guise of the Good, *in* Velleman, The Possibility of Practical Reason (1992).

[55] Augustine, The Confessions, supra note 54, at 36.

[56] Anscombe, Intention, supra note 9; Raz, Incommensurability, supra note 9; Raz, Agency, supra note 9.

[57] Cf. Judith Baker, Rationality Without Reasons, 117 Mind 763 (2008), on the difference between evaluating something and finding something appealing or desirable.

fact I must represent (if at all) by means of a further mental representation, a representation of drinking as good-insofar-as-it-is-pleasant. My more basic representation of drinking as pleasant is not in itself a representations of drinking as good. As against this, one might insist that pleasures, by their nature, always *seem* good (or, more radically, that taking pleasure in something just is a way of taking the thing to be good). But this position needs to be supported by an argument—an argument with premises beside the fact that all pleasures are appealing. To infer that all pleasures seem good from the fact that all pleasures are appealing is to beg the very question at issue, whether finding-something-appealing is the same as finding-it-good.

Now it is usually the case—perhaps it is always the case—that when we find something appealing, we find certain of its *features* appealing. This invites the thought that finding a thing appealing is just a matter of taking it to have certain appealing properties. But this thought is mistaken. Whether or not taking something to have certain appealing properties is necessary for psychic attraction, it is certainly not sufficient, because we can ascribe appealing properties to something without thereby being psychically attracted to it. Suppose that you have just entered a sauna and begun to sweat profusely. You know that within minutes you will have a dry sensation in your throat and that you will be able to alleviate the unpleasant sensation by drinking. Moreover, you are aware that, even now, before the dry sensation creeps into your throat, a sip of water would be pleasant and satisfying. Yet you are not currently thirsty: you have no unpleasant sensation in your throat and no felt motivation to drink. Does your mere awareness that drinking would be pleasant constitute a desire to drink? Of course not. At any given moment, you may be consciously aware of many possible pleasant experiences and vividly aware of what makes them pleasant, yet not desire any of them. You can be certain that sexual activity would be pleasurable and even represent this to yourself graphically, yet have no desire to engage in it. You can be certain that revenge would be delicious, yet feel no vengeful urge.[58]

[58] Cf. Alexander Nehamas, Only a Promise of Happiness: The Place of Beauty in a World of Art 68 (2007), on "the serious difference between describing a face as attractive and actually being attracted to it."

Another possibility is that psychic attraction consists in selective attention—specifically, selective attention to those very features on account of which you find something appealing. In this vein, consider Scanlon's notion of a desire in the directed-attention sense: "A person has a desire in the directed-attention sense that P if the thought of P keeps occurring to him or her in a favorable light, that is to say, if the person's attention is directed insistently toward considerations that present themselves as counting in favor of P."[59] Scanlon's notion of a desire in the directed-attention sense gets us closer to what it is to be psychically attracted to something, but not close enough. For we can easily imagine that as you sit in the sauna, your attention is directed insistently toward considerations that render drinking appealing. You cannot but think of how dehydrated you are becoming and of how pleasant it would be to take a sip of water. Yet your throat remains comfortable and you feel no attraction to drinking.

None of this is to deny that, when you desire something, your attention ordinarily is directed to appealing properties of the thing that you desire. My point is that the desire is what directs your attention; attention is not what constitutes your desire.

These reflections indicate that our being psychically attracted to something is not simply a matter of our ascribing appealing properties to the thing, even when we attend to these properties insistently. This seems especially true in cases of romantic or sexual desire. Ascription of appealing properties is obviously insufficient for romantic or sexual desire, as we can recognize that a person has appealing features (and even find these features utterly captivating) without being attracted to the person. What may be less obvious is that the ascription of appealing features is not only insufficient for romantic or sexual desire but also unnecessary. When we desire someone romantically, we are often unable to elucidate the basis of our attraction. Although we can usually specify personal attributes that entice us, being attracted to someone is different from being enticed by the person's attributes. We are attracted to the person as a whole, not to the person *qua* bearer of certain attributes. Even if we can ordinarily answer the question "what do you see in her?" by specifying personal attributes that we genuinely find attractive, these answers tend to miss the point. The same is true

[59] Scanlon, What We Owe, supra note 7, at 39 (emphasis added).

for sexual desire. If you are sexually attracted to someone, as opposed to merely judging that she is sexually attractive, that is not a matter of finding certain of her attributes or mannerisms sexually attractive. We are sexually attracted to *people*, not ordinarily to their features.

I've now considered and rejected four candidates for the role of psychic attraction: (i) felt motivation, (ii) evaluative 'seeming,' (iii) ascription of appealing properties to the desired object, and (iv) insistent attention to such properties. Each of these phenomena might ordinarily accompany psychic attraction. Yet none is sufficient for it; thus, none should be thought the essence of desire.

Even if the foregoing four phenomena (or some subset thereof) were somehow jointly sufficient for psychic attraction, I still would deny that they could *constitute* psychic attraction. As we saw when reflecting on romantic and sexual desire, psychic attraction can occur in the absence of all of these phenomena except felt motivation. But if psychic attraction can occur in the absence of things that are putatively sufficient for it, then these things cannot constitute psychic attraction, cannot be what it is.[60] So even if some subset of these four phenomena is sufficient for psychic attraction, we should continue to regard psychic attraction as something distinct from the subset—something the subset can cause, perhaps, but not something it can constitute.

I am not suggesting that psychic attraction could be isolated from *all* of its usual accompaniments—especially not felt motivation, which seems always to accompany desire. Yet felt motivation is never brute, never the bedrock of desire. On the contrary, we feel motivated to pursue the things we desire *because* those things psychically attract us. Far from being reducible to felt motivation, attraction is what grounds felt motivation and makes it intelligible. From the standpoint of introspection, it is our being psychically attracted to something that makes sense of our feeling motivated to pursue it, not the other way around.

As I've said, I suspect that the feeling of psychic attraction is itself indefinable, that it can't be analyzed in terms of more primitive psychic phenomena. I haven't shown this conclusively, of course, because I haven't (and couldn't have) considered every possible candidate for reduction. But I hope at least to have presented a coherent and

[60] I am assuming that psychic attraction is not multiply realizable at the level of conscious psychological states.

plausible alternative to the prevailing theories of desire, theories that mistake psychic attraction for one of its usual accompaniments. If I have succeeded, then any theory of desire must accommodate psychic attraction—either as analytically primitive or as reducible to other (herein unexamined) psychic phenomena.

## VI. Psychic Attraction and the Provocation Defense

Returning to the criminal law, I now consider whether the proposed theory of affective desire resonates harmoniously with the common-sense conception of human action that undergirds the volition-based criminal defenses. For demonstrative purposes, I focus on the defense of provocation.

What is going through your mind when you are provoked? Ordinarily, there is at least an insistent thought that the provoker has wronged you and deserves to suffer[61] as well as an intense motivation to strike back. Yet these two psychic states cannot be the whole story, because neither the thought nor the motivation can give rise to intelligible akratic action, even when they operate in tandem (see supra Part IV). There must be something more.

And there is. Beside the thought and the motivation, there is a *desire*—the desire to retaliate. I suggest that we understand this desire as an instance of psychic attraction. When you are provoked, you find the prospect of retaliation attractive. You are drawn to it. You find it appealing—even pleasurable. I mean this literally. (It was not a coincidence that the desires I scrutinized earlier were almost all desires for various kinds of pleasure.) Poets have long described revenge as "sweet"[62] and neuroscience suggests that this description is the literal

[61] The thought that the wrongdoer deserves to suffer need not be a thought that you endorse; in other words, the thought need not be a *belief*. This thought therefore can co-exist easily (if uncomfortably) with a belief to the effect that retaliation is unjustified—a belief that renders your subsequent act of retaliation akratic.

[62] E.g., Homer, The Iliad XVIII, 109 (Samuel Butler trans., The Internet Classics Archive), *available at http://classics.mit.edu/Homer/iliad.mb.txt*,

who describes anger in anticipation of revenge as "sweeter than drops of honey." See also John Milton, Paradise Lost IX, 171, *available at http://www.dartmouth.edu/~milton/reading_room/contents/index.shtml*; Lord Byron, Don Juan I, 124, *available at http://www.gutenberg.org/files/21700/21700-h/21700-h.htm*.

truth, or nearly so. When a victim of unfair treatment inflicts so-called altruistic punishment[63] on someone who has wronged him,[64] the victim undergoes brain activity associated with the anticipation of such things as the pleasant taste of sugar.[65] It is not merely that the victim experiences acts of retaliation as pleasurable; he is *attracted* to the prospect of retaliation in the same way as he would be attracted to the prospect of eating sugar. He is drawn to retaliation as though it were "sweeter than drops of honey."[66] I take this neuroscientific finding to be consistent with the phenomenology of retaliation. Amid all of the psychic turmoil caused by an outrageous provocation, there is a powerful current of psychic attraction. We should not assume that this sort of pleasure-directed attraction can occur only in happy times.

If psychic attraction really is the essence of retaliatory temptation, can retaliatory temptation (so conceived) really give rise to intelligible akratic action? I see no reason why it cannot, as I see no reason to deny that a person can be attracted *most* to something other than what he judges best. Psychic attraction is one thing; judgment is another. Regrettably, these two things do not always go hand in hand.

You may find this answer wholly unsatisfying. You may feel that no matter how we conceive of desire, akratic action will remain as mysterious as ever. So it may. But to solve this mystery—if that is really what it is—we need more than a conception of desire. We need a conception of agency. A conception of desire cannot be expected to solve the mystery of akratic action by itself. All we can expect of such a conception is that it not render the mystery of akratic action insoluble. Akratic action may yet be ineluctably mysterious, but that is not because the nature of desire makes it so. The leading theories of temptation fail this basic requirement, rendering akratic action mysterious or impossible.

[63] Altruistic punishment is defined as punishment inflicted on a perpetrator by a victim, at some cost to the victim. See Dominique J.-F. de Quervain, et al, The Neural Basis of Altruistic Punishment, 305 Science 1254 (2004).

[64] All the subjects in the study described in de Quervain, The Neural Basis of Altruistic Punishment, supra note 63, were male.

[65] De Quervain, The Neural Basis of Altruistic Punishment, supra note 63; John P. O'Doherty, et al, Neural Responses during Anticipation of a Primary Taste Reward, 33 Neuron 815 *(*2002). See also Brian Knutson, Sweet Revenge?, 305 Science 1246 (2004).

[66] Homer, The Iliad, supra note 62, at XVIII, 109.

Recall the difficulties that beset the evaluative conception. First, there is the problematic account that evaluative theorists typically give of akratic action—the account that proposes an analogy between, on the one hand, evaluative appearance and action and, on the other, perceptual illusion and belief. This account ultimately fails to make sense of akratic action, I explained, because it fails to make sense of how we could be led astray by evaluative illusions that we recognize as illusions. As I argued, the standard evaluative account is actually worse than unhelpful; it is fatal. Instead of showing how akratic action is intelligible, the standard evaluative account seems to entail that akratic action is impossible. If a known perceptual illusion cannot tempt or cause belief, then—taking seriously the analogy between evaluative appearance and perceptual illusion—it would seem to follow that a known illusory evaluative appearance cannot tempt or cause action. Moreover, even if a perceptual illusion could *tempt* belief, I observed, nothing could be easier than resisting such a doxastic (i.e., belief-related) temptation; and if nothing could be easier than resisting a doxastic temptation, then—again, taking seriously the analogy between evaluative appearance and perceptual illusion—it would seem to follow, in violation of the first truism, that nothing could be easier than resisting a *practical* temptation.

Stepping back from the standard evaluative account of akratic action, I then argued that, on reflection, evaluative appearance simply seems incapable of exerting any influence on action. It is mysterious how I could be motivated by an *evaluative* illusion (that the whipped cream's deliciousness illusorily appears good) yet not be motivated by a *non-evaluative* illusion (that the imitation whipped cream illusorily appears delicious). As I observed, if an evaluative theorist were to respond by insisting that evaluative illusion simply possesses a kind of motivational efficacy that non-evaluative illusion lacks, she would incur a heavy burden: to explain why evaluative illusion possesses motivational efficacy yet lacks *doxastic* efficacy, in other words, to explain how evaluative illusion can make us *do* something without at the same time making us *believe* something. What mechanism, I asked, could persuade us to act as though the world were a certain way without thereby persuading us to represent the world as being that way? If this question lacks a plausible answer—and it seems to—then the

evaluative theorist's response entails that any temptation capable of motivating us to act against our all-things-considered judgment would also be capable of getting us to revise that judgment, thereby rendering our subsequent conduct non-akratic.

None of these problems plagues the idea that affective desire consists in psychic attraction because psychic attraction is fundamentally non-representational. Feeling attracted to an object is not a matter of representing the object as *being* some way or other. Much less is it a matter of representing the object as being good. How a person represents some object is a separate question from whether the object attracts him. So if we conceive of temptation in terms of psychic attraction, we will avoid the seemingly insurmountable challenge of explaining how a person could be led astray by a representation the content of which he does not endorse. In other words, we will avoid the problem of explaining how a known illusion could somehow be hard to ignore. Thus, there will be no barrier to our accepting the truism that resisting temptation is ordinarily difficult. Nor will there be any barrier to our retaining the host of volition-based criminal defenses that this truism undergirds.

Can conceiving of temptation in terms of psychic attraction give us any insight into the first truism? Can it help us understand why resisting temptation is ordinarily difficult? I am not sure that it can. Nor am I sure that that is a problem. Perhaps we must simply accept the truism as bedrock, as something that defines the character of attraction. We were bound to hit bedrock eventually, and this seems as good a place to stop as any other.

Next, recall the difficulties that beset the mechanistic conception of temptation. Like the evaluative conception, the mechanistic conception renders akratic action seemingly impossible. But it does so for a different reason. Here, the apparently insurmountable challenge is not the challenge of explaining how resisting temptation might be difficult—the mechanistic conception has no trouble with that. Rather, it is the challenge of explaining how a person who is not subject to an irresistible compulsion could, in defiance of his all-things-considered judgment, intelligibly pursue what he does not find in any way appealing. This challenge arises because the mechanistic conception takes temptation to be nothing but a vector of "sheer brute force," a raw motivating state that does nothing to present its object as appealing

or attractive. If I were afflicted by such an impulse and it were not irresistible, what could make sense of my acquiescing to it? Consistent with the mechanistic conception, we could not make sense of my acquiescing by appealing to the pleasure of acquiescence, because that pleasure could make sense of my acquiescing only if we understood my attraction to the pleasure in non-mechanistic terms. If we understood the attraction mechanistically, we would push the problem back one level rather than solve it. For the same reason, we could not make sense of my acquiescing to a resistible brute urge by appealing to the displeasure of resistance, because that displeasure likewise could make sense of my acquiescing only if we understood my aversion to the displeasure in non-mechanistic terms. On pain of rendering akratic action unintelligible, then, the mechanistic conception must contravene the second truism about temptation and suppose temptation to be irresistible, making provocation and duress seem like they should be complete defenses instead of partial ones.

If we conceive of temptation in terms of psychic attraction, there will be no barrier to our accepting that resisting temptation is not ordinarily impossible. Temptation *qua* psychic attraction need not be irresistible in order for our acquiescence to be intelligible. As I said, it is often the case that we succumb to temptation not because resistance would be futile but because it would be no fun. (Recall the example of the person who finds himself on the verge of committing a sexual indiscretion that he could avoid without much effort.) The idea that temptation consists in psychic attraction therefore is wholly consistent with the second truism.

But can the notion of psychic attraction *explain* this truism? Can it help us understand why the truism is true? As before, I am not sure that it can, nor am I sure that that is a problem. In all likelihood, we must accept the truism as bedrock, as something that defines the character of attraction.

## Conclusion

The criminal law's volition-based defenses rest on two psychological assumptions, assumptions that are inconsistent with the leading theories of temptation but harmonious with the proposed theory of affec-

tive desire: that resisting a powerful impulse is extremely difficult; and, at the same time, that no matter how difficult it might be to resist a powerful impulse, doing so is not ordinarily impossible. We should be loath to abandon these assumptions. They accord with our deepest intuitions about human nature. They also inhere in the criminal law, an institution that derives its content not merely from a priori reflection but also, more fundamentally, from the cumulative wisdom of society's past efforts to regulate conduct and dispense just punishment. Taking the criminal law seriously will show us what temptation cannot be, as well as what it could be.

## Acknowledgments

I received helpful feedback on the material in this essay from many people, including Bruce Ackerman, Paul Audi, Robert Audi, Sarah Buss, Steve Darwall, Chris Essert, David Estlund, Tamar Gendler, Liz Harman, Verity Harte, Scott Hershovitz, Shelly Kagan, Colin Klein, Adrienne Lapidos, Jed Lewinsohn, Alexander Nehamas, Japa Pallikkathayil, Robert Post, Peter Railton, Don Regan, Michael Della Rocca, Gideon Rosen, Scott Shapiro, Matt Smith, Michael Smith, Zoltan Szabo, David Velleman, Bruno Whittle, Ken Winkler, and Kevin Zaragoza.