Published in *The Kant Yearbook* 4, 1 (2012)133-142

Kant's Two Solutions to the Free Rider Problem[1]

Adrian M. S. Piper

Instrumentalist justifications of natural law, beginning with Hobbes' Social Contract theory, usually rely on a hypothetical narrative in which self-interestedly rational agents agree to exchange their unlimited freedom in the state of nature for the peace and stability attendant on abiding by certain rules held in common. The Social Contract then consists in their consensual promise to one another to abide by these rules, even when doing so conflicts with their immediate self-interest. In return, they all receive the long-term benefits of laws governing private property, contract, freedom of speech and the like; as well as of moral conventions such as honesty, reliability, and charity. Rawls was the first to formalize this idea game-theoretically in his *Theory of Justice*. Recent game-theoretic accounts of the origins of interpersonal coordination[2] take their cue from Rawls' formulation.

An Instrumentalist justification of natural law naturally engenders the Free Rider problem, because this problem arises by carrying the Instrumentalist justification to its logical extreme. Hobbes' Foole[3] was the first to reason that if self-interest justifies exchanging the state of nature for the Social Contract, then self-interest also justifies violating the Social Contract for personal gain. The Free Rider ostensibly promises to obey the rules as consensually agreed, with the intention of breaking that promise when this is personally advantageous. She exploits others' renunciation of immediate self-interest in order to advance her own. If all agents reason similarly, no Social Contract is possible. So self-interest does not justify a Social Contract, and individually self-interested rationality would seem to be collectively self-defeating.

---

[1] © APRA Foundation Berlin 2012. This essay is excerpted from a longer discussion, *Kant's Metaethics: First Critique Foundations* (in progress). An earlier version was delivered to the first plenary session of the United Kingdom Kant Conference, *Reading Kant*, at the University of St. Andrews in September 2011 under the title, "Kant's Two Replies to Hobbes." I am grateful for comments from the audience, and most particularly from Sorin Baiasu, Martin Sticker, and Jens Timmermann. Comments and criticisms from an anonymous referee for the *Kant Yearbook* have much improved the final draft.

[2] See, for example, David Lewis, *Convention: A Philosophical Study* (Cambridge, Mass.: Harvard University Press, 1969); Allan Gibbard, "Utilitarianisms and Coordinations" (Ph.D. diss., Harvard University, 1971); Edna Ullman-Margalit, *The Emergence of Norms* (Oxford: Clarendon Press, 1977). Lewis' book predates the publication of Rawls' *Theory of Justice*, and Gibbard completed his dissertation in the same year. However, both were graduate students at Harvard while Rawls was teaching and circulating his book in manuscript form.

[3] Cf. Thomas Hobbes, *Leviathan*, Ed. Michael Oakeshott (New York: Macmillan/Collier Books, 1977), 115-117.

Kant was well aware of the Free Rider problem posed by Hobbes' Foole. Kant regarded it as a by-product of Hobbes' deficient conception of reason, which exempts self-interest from the cognitive functions of generalization and universalization. For Hobbes, these functions are of merely instrumental value, in grasping the external causal relations and principles through which we satisfy our desires. For Kant, by contrast, they are necessary conditions for unified experience of any kind, including experience of those desires themselves. If a perceived self-interest must meet the same rationally rule-governed cognitive requirements as any other perceived state of affairs, then there is no basis for exempting it from the rationally rule-governed cognitive requirements of moral principle in particular. This leaves the Free Rider with no justification for violating the Social Contract at all, not even a self-interested one.

This is the conception of reason behind the two successive and interconnected solutions to the Free Rider problem that Kant offers in the first *Critique* and *Groundwork*. He of course did not have the contemporary concept of a Free Rider per se. Rather, he identifies what are in fact Free Riders as the most noxious species of *polemicists*, for whom he reserves a special place in hell. Polemicists, for Kant, are those who attack metaphysical beliefs in the existence of God, freedom or immortality by harping on their lack of empirical proof, in order to buttress and conceal the equally flimsy metaphysical foundations of their own cynicism. Kant thinks polemic debases the stature and authority of reason, reducing it to a method of squabbling that destabilizes social equilibrium and portends disintegration into the Hobbesian state of nature. He is particularly enraged by the use of this tactic to pseudorationalize[4] our moral derelictions, which only serves to accelerate our downward slide. In the first *Critique,* Kant agrees with Hobbes that this process of deterioration can only be reversed through the consensual agreement to relinquish the unlimited freedom of the state of nature, and submit to the authority of law.

To secure that agreement, Kant proposes two separate but textually related solutions to the Free Rider problem: First, a critique of reason in its polemical use. Kant articulates this proposal in the first *Critique* – and, in the *Groundwork*, applies it to the Free Rider's self-defensive polemical subterfuge. Second, he argues that promise-keeping is a perfect duty that allows no exceptions "to the advantage of inclination." These two solutions appear as connected steps in Kant's attempted derivation of perfect and imperfect

[4] I develop this concept at length in Chapter VII. "Pseudorationality," of my *Rationality and the Structure of the Self, Volume II: A Kantian Conception* (http://www.adrianpiper.com/rss/docs/Rationality%20and%20the%20Structure%20of%20the%20Self,%20Volume%20II-%20A%20Kantian%20Conception.pdf , 2008), 254-278.

© Adrian Piper Research Archive Foundation Berlin

duties from the categorical imperative. The questionable success of the derivation does not affect the independent merit of either solution. The first enables us to better appreciate the role of those laws in structuring and regulating our empirical agency. The second enables us to mend the Social Contract and reverse our descent into Hobbes' state of nature. The first solution enables us to see the point of the second.

1. Polemical "Reasoning" and the Free Rider Mentality

In the first *Critique*, Kant formulates the Free Rider problem as one of coordinating among agents who have conflicting beliefs and agendas, but who also have the option of reconciling their disputes through appeal to rational rules whose governing authority is not in doubt. Each such agent must choose between constraining his claims by following those rules and thus furthering stability for all; or else violating them for personal gain and thus edging everyone closer to social disorder:

> (1) (1) One can regard the critique of pure reason as the true court of law for all such disputes, (2) for it is not involved in these disputes, (3) which as such are immediately concerned with objects; (4) but rather is oriented toward determining and judging the scope of entitlement [*Rechtsame*] of reason in general, (5) according to the principles of its first institution.
>
> (6) Without this, reason is, as it were, in the state of nature, (7) and can only validate and secure its claims and demands through *war*. (8) By contrast, the critique that draws all of its decisions from the fundamental rules of its own formation, (9) whose stature [*Ansehen*] no one can doubt, (10) provides us the calm of a lawful condition, (11) in which we are to conduct our disputes in no other way than through the *legal process*. (12) What the negotiation concludes in the first state is a *victory*, (13) of which both sides boast, (14) and upon which a merely insecure peace follows that is granted by a mediating authority; (15) in the second state, however, (16) it concludes a *judicial sentence*, (17) which, because it affects the very source of the disputes themselves, (18) must bestow an eternal peace. (19) Thus the endless disputes of a merely dogmatic reason finally demand the search for calm in some kind of critique of reason itself, (20) and in a legislation based upon it. (21) As Hobbes claims, the state of nature is a state of injustice and violence, (22) and we are forced to abandon it (23) in order to submit ourselves to legal constraints (24) that limit our freedom (25) so that it can be consistent with others' freedom (26) and the common good of all (KrV, A 751.23-40; A 752.01-14).[5]

[5] (1) Man kann die Kritik der reinen Vernunft als den wahren Gerichtshof für alle Streitigkeiten derselben ansehen; denn sie ist in die letzteren, als welche auf Objekte unmittelbar gehen, nicht mit verwickelt, sondern ist dazu gesetzt, die Rechtsame der

© Adrian Piper Research Archive Foundation Berlin

Passage (1) occurs in Part II of the *Critique of Pure Reason*, the Transcendental Doctrine of Method, in Chapter I, "The Discipline of Pure Reason." By a *discipline*, Kant means the force through which a constant tendency to deviate from certain rules is restrained and finally eliminated, and a present habit thereby extinguished (KrV, A 710. 06-12). So Part II, Chapter I treats of our need to restrain reason from transcending the legitimate boundaries set by experience, and to replace the habit of untrammeled metaphysical speculation with rational circumspection. Passage (1) is to be found in Section II of that chapter, entitled "The Discipline of Pure Reason in respect of its Polemical Employment." By the *polemical* employment of pure reason, Kant means the use of reason to defend a claim, not by denying that it may be false, but rather by arguing that its opposite cannot be proven to be true (KrV, A 739.22-27, and A 740.01-06).

Kant is referring specifically to the metaphysical disputes about the existence of God, freedom and immortality that he has shown in Part I of the first *Critique* to lead to antinomies, because of the failure of these disputes to respect reason's limitations. In the Dialectic, he contended that the antinomies could be resolved by distinguishing between appearances and things in themselves, hence that the disputes in question resulted from misunderstanding. Those well-intentioned but misguided attempts to reason through to a conclusive answer to these questions are not Kant's target here. Instead he is criticizing the *polemical method* of winning these disputes: negatively attacking an opposing view, in order to buttress the credibility and conceal the flimsy foundations of one's own.

Vernunft überhaupt nach den Grundsätzen ihrer ersten Institution zu bestimmen und zu beurteilen.

Ohne dieselbe ist die Vernunft gleichsam im Stande der Natur, und kann ihren Behauptungen und Ansprüche nicht anders geltend machen, order sichern, als durch *Krieg*. Die Kritik dagegen, welche alle Entscheidungen aus den Grundregelen ihrer eigenen Einsetzung hernimmt, deren Ansehen keiner bezweifeln kann, verschafft uns die Ruhe eines gesetzlichen Zustandes, in welchem wir unsere Streitigkeit nicht anders führen sollen, als durch *Prozeß*. Was die Händel in dem ersten Zustande endigt, ist ein *Sieg*, dessen sich beide Teile rühmen, auf den mehrenteils ein nur unsicherer Friede folgt, den die Obrigkeit stiftet, welche sich ins Mittel legt, im zweiten aber die *Sentenz*, die, weil sie hier die Quelle der Streitigkeiten selbst trifft, einen ewigen Frieden gewähren muß. Auch nötigen die endlosen Streitigkeiten einer bloß dogmatischen Vernunft, endlich in irgendeiner Kritik dieser Vernunft selbst, und in einer Gesetzgebung, die sich auf sie gründet, Ruhe zu suchen; so wie Hobbes behauptet: der Stand der Natur sei ein Stand des Unrechts und der Gewalttätigkeit, und man müsse ihn notwendig verlassen, um sich dem gesetzlichen Zwange zu unterwerfen, der allein unsere Freiheit dahin einschränkt, daß sie mit jedes anderen Freiheit und eben dadurch mit dem gemeinen Besten zusammen bestehen könne. (KrV, A 751.30-40 and A 752.01-14).

© Adrian Piper Research Archive Foundation Berlin

Polemical reasoning conducts the disputes about the existence of God, freedom and immortality entirely within the inaccessible realm of things in themselves, where it is, in fact, impossible to conclusively prove any claim to be either true or false, because neither the opposing view nor one's own is anchored in empirical experience. The disputes are indeed about objects (clause 1.3), but not objects of the senses. Because we lack experiential access to such objects, our arguments for or against their existence consist in unfounded inferences. From this state of affairs, Kant concludes that since reason cannot and never will be able to conclusively *dis*prove the existence of God, freedom or immortality, we may confidently continue to believe in them, on grounds of their practical value. And we may leave authentic freethinkers such as Hume and Priestley to exercise their faculties of reason and skepticism on these metaphysical questions, and to discover in good faith the limits of dialectical debate (KrV, A 747.17-23).

However, Kant distinguishes such authentic freethinkers from polemicists, for whom he has only dripping contempt. He accuses them of dishonesty (KrV, A 747.28, A 748.25), baseness (KrV, A 748.20), dissimulation (KrV, A 748.25-26, A 749.28), hypocrisy (KrV, A 748.26), personal vanity (KrV, A 749.07), duplicity (KrV, A 749.28), and fraud (KrV, A 749.28). The polemical method of reason, he charges, takes advantage of the weakness inherent in all such dogmatic metaphysical positions from a position of weakness itself (KrV, A 751.08-10). It defends an issue of fairness unfairly, and therefore ought not to exist (KrV, A 750.13-19). Polemics is a form of intellectual corruption so antithetical to reason itself, and such a futile exercise in posturing and shadowboxing, that in the end, Kant argues, a polemic in the field of pure reason in fact does not exist (KrV, A 756.11-16).

It is in this context that Kant offers the argument in passage (1); and it is to "such disputes" that he is referring in clause (1.1). He regards polemical reasoning as a last-ditch, no-holds-barred battle among conflicting agendas of intellectual self-aggrandizement, marked by underhanded and deceitful tactics; and the critique of pure reason he has offered as its antidote. But passage (1) does not imply that the publication of the *Critique of Pure Reason* will now silence all further polemics, liberating us to move on to the "calm of a lawful condition (1.10)." On the contrary: the disputes of a "merely dogmatic reason" are "endless," and therefore *always* "finally demand the search for calm in some kind of critique of reason itself (1.19)." Because reason's habit of metaphysical speculation is so difficult to discipline, the critique of polemical reasoning must be repeatedly administered.

It is tempting to suppose that both the polemical battle and the judicial conciliation promised by "some kind of critique of reason itself" take place only within the rarified confines of speculative philosophy, among scholars trained in its subtleties. This would be a mistake. The question of whether we are free or causally determined arises whenever we attempt to excuse moral

© Adrian Piper Research Archive Foundation Berlin

wrongdoing, whether our own or another's, on the grounds that the agent had no choice – that is, virtually every time someone commits a wrong. The question of whether or not God exists arises whenever wrongdoers question whether or not their instinctive fear of divine retribution is justified. The question of whether or not our souls are immortal arises whenever a wrongdoer wonders whether or not the present benefits of her wrongdoing might engender a future backlash that extends beyond the life in which she enjoys them; or whether the present punishment she endures implies some distant future redemption. Whenever wrongdoers attempt to evade moral responsibility, to ridicule their instinctive fear of God's wrath, or to belittle their anticipation of punishment, they depend on polemical reasoning to buttress their belief in their own moral impunity. Hence Kant's argument in passage (1) applies to the use of such polemics whenever they are used to defend wrongdoing itself.

This is one sense in which polemic without rational critique abandons reason to the state of nature (1.6), in which war (1.7), injustice and violence (1.21) are necessary in order to secure its claims. Kant's virtual state of nature consists not merely in battling metaphysicians, huffing and puffing and bluffing their way to victory in debate. Nor does it consist simply in battling human agents, committing acts of violence, treachery and disorder toward one another. Rather, Kant's virtual state of nature consists in human agents committing these acts *and* huffing and puffing and bluffing *themselves* in order to justify them; deploying that particularly cynical form of bad-faith self-defense in which worries about moral responsibility, far-reaching consequences and the condemnation of the universe for unconscionable acts are belittled or denied. Polemic clears a path for the fatalistic, anything goes reasoning that enables wrongdoers to believe they are beyond reproach and beyond the reach of the law. This is the essence of the Free Rider mentality, the opportunistic mindset that feeds on disordered social conditions in order to maximize occasions for self-seeking, while minimizing the obligations of promise-keeping and self-regulation. A critique of pure reason must "determin[e] and judg[e] the scope of entitlement of reason (1.4) according to the principles of its first institution (1.5)," because by reviewing those fundamental principles and delineating their outer limits of legitimate application, this critique effectively reveals polemical exertions as the empty wheel-spinning they really are.

So Kant's contention that a critique of pure reason is needed in order to transcend the state of nature (1.12-18) should not be supposed to refer only to settling the philosophical disputes of trigger-happy speculative metaphysicians who stoop to polemics in order to shore up an untenable position. Rather, he means to propose a critique of pure reason as a more general method for literally bringing human agents in a disordered or unstable social state *to their senses*. By analyzing reason's cognitive

© Adrian Piper Research Archive Foundation Berlin

foundations "according to the principles of its first institution (1.5)," reminding us of its proper range of operation, and demarcating the outer limits of its authority, such a critique calls attention to the defining function of reason in the law-governed organization of the self. It returns us from the futile shadowboxing of polemical attacks on moral responsibility (KrV, A 756.11-16) to concrete empirical recognition of it in the self-regulation we habitually exercise. Passage (1) thus offers a perfectly general proposal for resolving both the diverse and conflicting agendas of undisciplined individuals in an unstable social state, and also their flawed and conflicting strategic reasoning about how to realize those agendas.

Kant's proposal is that this reasoning itself must be criticized, independently of the dogmatic strategies it serves, by reference to the "fundamental rules of [reason's] own formation (1.8), whose stature no one can doubt (1.9)." The fundamental rules of reason's formation are the rules of transcendental logic laid out in the first *Critique's* Table of Judgments and Table of Categories. Their stature and authority are so indubitable that coherent experience itself would be impossible without them. Rational legislation – and self-legislation – must be based on these universal rules, not on the individual agendas or ends in whose service we instrumentalize them. The Free Rider's mistake is, in effect, a failure of vision; a failure to situate and evaluate his individual agendas and ends relative to the rational principles that make them possible. Restoring a sense of perspective to our choices requires us to detach our processes of reasoning from the warring ends and interests they are intended to promote (1.2-3); and evaluate that reasoning itself according to the universal rules of reason to which it aspires (1.8). Kant maintains that only critically reflective decision-making that obeys these fundamental rules, irrespective of the diverse uses to which they are put by "merely dogmatic reason (1.19)," "provides us the peace of a lawful condition (1.10), in which we are to conduct our disputes in no other way than through the *legal process* (1.11)."

2. Kant's First Solution: The Critique of Reason

In the *Groundwork of the Metaphysic of Morals*, Kant provides just such a critique of the Free Rider's faulty strategic reasoning, from the distanced perspective of the "fundamental rules of [reason's] own formation (1.8)." He directly applies passage (1)'s argument to the polemical gyrations by which we attempt to justify moral wrongdoing, when this is parasitic upon a "merely insecure peace … granted by a mediating authority (1.14)" such as the sovereign body. These are the actual circumstances – not those of a lawless state of nature as Hobbes originally described it – under which we normally violate the moral law. At least on the face of it, the problem for us and for the sovereign body alike is not how to escape the state of nature. It is rather how to prevent ourselves from slipping back into it, through our

© Adrian Piper Research Archive Foundation Berlin

repeated, self-serving violations of the moral law whose authoritative stature to guide our actions (1.9) we have already acknowledged.

But Hobbes' own solution to this threat to the stability of the Social Contract – that the Foole's declaration of her intentions to violate it for personal gain would draw upon her the punitive retribution of the sovereign body and the ostracism of her fellow citizens – was clearly inadequate: a rational Foole will break the rules secretly.[6] Kant, by contrast, rightly rethinks Hobbes' Foole as a game-theoretically sophisticated Free Rider, who arrogates to herself the liberty to break pre-existing law for personal gain provided that others continue to obey it. The stability of the Social Contract is undermined, not by the renegade who publicly declares his criminality, but rather by the hypocrite who publicly upholds the Social Contract while privately violating it.

The challenge, as Kant sees it, is then to *replace* the resulting "merely insecure peace" by a final verdict from the highest court of appeals, "which, because it affects the very source of the disputes themselves (1.17), must bestow an eternal peace (1.18)." The highest court of appeals is not the sovereign body, but rather the set of foundational principles that define reason itself. The insecurity of the peace under which we and the sovereign suffer equally is due to the temptation to free ride on basic rational laws to which we have already consented, and to pseudorationalize this by regressing back to the "endless disputes of a dogmatic reason (1.19)." This poisons both the security of that peace, and also our respect for the laws we have established.

Passage (2) analyzes with a touch of sarcasm the labyrinthine tinkering of the Free Rider's polemical self-justification:

> (2) (1) If we now attend to ourselves whenever we transgress a duty, (2) we find that we do not really will, [such that] our maxim should become a universal law, (3) because that is impossible for us; (4) but rather the opposite is itself in reality to remain universally a law. (5) We only take the liberty of making an *exception* to it for ourselves (or only just for this once) (6) to the advantage of our inclination. (7) Consequently, if we were to consider everything from one from one and the same standpoint, namely reason, (8) we would come across a contradiction in our own will, (9) namely that a particular principle should be objectively necessary as universal law, (10) and yet subjectively not be universally valid, but rather should allow exceptions. (11) But as we first consider our action from the standpoint of a will wholly in accord with reason, (12) but then exactly the same action from the standpoint of a will affected by

[6] This is the historical reference behind the title of Sen's critique of revealed preference theory. See his "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory," *Philosophy and Public Affairs 6*, 4 (1977), 317-44.

© Adrian Piper Research Archive Foundation Berlin

> inclination, (13) actually there is no contradiction here, (14) but rather an opposition of inclination to the prescription of reason (*antagonismus*), (15) through which the universality of the principle (*universalitas*) is turned into a mere generality (*generalitas*), (16) in such a way that the practical principle of reason is supposed to [*soll*] join up halfway with the maxim (GMS Ak. 04:424.18-39).[7]

Clause (2.1) indicates Kant's intent to offer an analysis of any and all transgressions of moral duty. Clause (2.4) implies that all of them occur against a background of perceived general compliance with the moral law. On the other hand, clause (2.6) uses exactly the same words and idiomatic expression [in boldface below] as he did earlier, when he defined a perfect duty as one that "permits no exception to the advantage of inclination (GMS Ak. 04:421.27-28 and fn.)":

> (3) … die keine **Ausnahme zum Vorteil der Neigung** verstattet (GMS Ak. 04:421.28, fn.05) ... .[8]
>
> (4) (2.6) … für uns ... **zum Vorteil unserer Neigung davon eine *Ausnahme*** zu machen (GMS Ak. 04:424.23-24).

The conjunction of (3) and (4) may suggest that (2.6) refers to violations of perfect duties in particular. However, clauses (2.1) and (2.6) are consistent under the assumption that (2.6) refers to the violation of any duty, whether perfect or imperfect, that the agent recognizes as requiring fulfillment at that

---

[7] (2) Wenn wir nun auf uns selbst bei jeder Übertretung einer Pflicht Acht haben, so finden wir, daß wir wirklich nicht wollen, es solle unsere Maxime ein allgemeines Gesetz werden, denn das ist uns unmöglich, sondern das Gegenteil derselben soll vielmehr allgemein ein Gesetz bleiben; nur nehmen wir uns die Freiheit, für uns (oder auch nur für diesesmal) zum Vorteil unserer Neigung davon eine *Ausnahme* zu machen. Folglich, wenn wir alles aus einem und demselben Gesichtspunkte, nämlich der Vernunft, erwögen, so würden wir einen Widerspruch in unserem eigenen Willen antreffen, nämlich daß ein gewisses Prinzip objektiv als allgemeines Gesetz notwendig sei und doch subjektive nicht allgemein gelten, sondern Ausnahmen verstatten sollte. Da wir aber einmal unsere Handlung aus dem Gesichtspunkte eines ganz der Vernunft gemäßen, dann aber auch ebendieselbe Handlung aus dem Gesichtspunkte eines durch Neigung affizierten Willens betrachten, so ist wirklich hier kein Widerspruch, wohl aber ein Widerstand der Neigung gegen die Vorschrift der Vernunft (*antagonismus*), wodurch die Allgemeinheit des Prinzips (*universalitas*) in eine bloße Gemeingültigkeit (*generalitas*) verwandelt wird, dadurch das praktische Vernunftprinzip mit der Maxime auf dem halben Wege zusammenkommen soll (GMS Ak. 04:424.18-39).

[8] The full sentence runs as follows:

Übrigens verstehe ich hier unter einer vollkommenen Pflicht diejenige, **die keine Ausnahme zum Vorteil der Neigung** verstattet, und da habe ich nicht bloß äußere, sondern auch innere vollkommene Pflichten, welches dem in Schulen angenommenen Wortgebrauch zuwiderläuft, ich aber hier nicht zu verantworten gemeint bin, weil es zu meiner Absicht einerlei ist, ob man es mir einräumt oder nicht (GMS Ak. 04:421.28, fn.04-10).

© Adrian Piper Research Archive Foundation Berlin

moment. Kant's definition of a perfect duty in the footnote to Ak. 421 implies that an *im*perfect duty, by contrast, does permit exceptions to the advantage of inclination. Yet once an agent has determined that this particular situation – for example, in which one is morally obligated to help one's elderly fellow pedestrian across the street – permits no such exception, she may still nevertheless try to exempt herself by invoking the faulty reasoning that passage (2) describes.

Clause (2.6) is also ambiguous with regard to its scope: That which is "to the advantage of our inclination" may extend only to exempting ourselves from the moral law (2.5); or it may extend as well to transgressing the law in the first place, and both trying and failing to represent that transgression as itself a law [(2.1), (2.2), (2.3) and (2.4)]. Assigning (2.6) the narrower scope would exclude the natural explanation of why we transgress our duty, namely self-interest; and why we cannot "really will (2.2)" that all others transgress their duty, too. The natural explanation is that both conditions are, in fact, "to the advantage of our inclination." Assigning the wider scope to (2.6) supports the observation that we can only successfully indulge our own transgressive inclinations if others do not simultaneously indulge theirs.

Thus Kant's analysis in passage (2) illuminates the Free Rider's vacillations, self-contradictions, and cyclical reversals, which heed the imperatives of rational principle one minute, and the impulse to self-seeking that interferes with it the next. The Free Rider transgresses the law when this serves his self-interest (2.1) and half-heartedly construes his transgression itself in lawlike terms (2.2). But he also tries to retain the law he has transgressed as "universally a law (2.4)," in order to enjoy the benefits of shared obedience to it. The Free Rider wants the liberty to break pre-existing rules for personal gain, provided that others continue to obey them. He wants the advantages and security of others' compliance with the rule in question, precisely in order to "take the liberty of making an exception to it for [himself] (or only just for this once) (2.5), to the advantage of [his] inclination (2.6)." So long as the Free Rider aspires to rational action at all, his attempt to rationalize his derelict intention induces in himself a contradiction in his own will, in which his derelict action should be universalizable, yet should not; should be an exception to the rule, yet itself the rule; should be rationally justifiable, yet not rationally conceptualized. His irrationality consists in contradicting himself, in being at war with him*self*, and not merely with his intellect.

Passage (2) is not the first in which Kant has addressed the Free Rider's faulty exceptionalist reasoning. He has already said, about the man contemplating whether or not to neglect the cultivation of his natural gifts,

© Adrian Piper Research Archive Foundation Berlin

> (5) He sees here, that a system of nature could always indeed exist under such a universal law …; only he cannot possibly *will* that this become a universal law of nature … (GMS Ak. 04:423.09-16),[9]

and, about the man deliberating about whether or not to help others in need,

> (6) But although it is possible that a universal law of nature could exist according to this maxim, it is nevertheless impossible to *will* that such a principle should hold everywhere as a law of nature (GMS Ak. 04:423.33-37).[10]

Kant's argument has been exactly the same in both cases. Someone who intends to transgress the law through passive neglect, whether of self or others, seeks to justify her action by willing it as itself a law that prescribes passive neglect, respectively of self (passage (5)) or of others (passage (6)). There is no conceptual inconsistency in this. However, such a will would, nevertheless, "be in conflict with itself, since many cases can arise in which [s]he needs … others' … assistance (GMS Ak. 04:423.37-38)."[11] Hence she also "does not really will" this law of passive neglect (2.2), because in both cases, she knows she can indulge her own passive neglect only if others do not indulge theirs. She intends, rather, that "the opposite is itself in reality to remain universally a law (2.4)," so as to continue to enjoy the advantages of others' conformity to it.[12]

In this internal conflict, the Free Rider takes the polemical standpoint of the "will affected by inclination (2.12)." Empirical inclination is strong, present and vivid; and this makes the intelligible commands of reason weak, faint and remote. The Free Rider takes advantage of the empirically concrete and particular character of felt inclination in order to reject the abstract, theoretical presumption that a particular moral principle is "objectively necessary as universal law (2.9):" The Free Rider is impelled by the belief that *this* drive, *this* desire, *this* derelict impulse at *this* moment must constitute a legitimate exception to the law, precisely because of its indexical strength, vividness, and presence; i.e. because of its concrete particularity right here and now. His inclination thus sets him in "opposition … to the prescription of reason (*antagonismus*) (2.14)," and confers on him the bogus authority to

---

[9] Da sieht er nun, daß zwar eine Natur nach einem solchen allgemeinen Gesetze immer noch bestehen könne, ...; allein er kann unmöglich *wollen*, daß dieses ein allgemeines Naturgesetz werde ... (GMS Ak. 04:423.09-16).

[10] Aber obgleich es möglich ist, daß nach jener Maxime ein allgemeines Naturgesetz wohl bestehen könnte, so ist doch unmöglich, zu *wollen*, daß ein solches Prinzip als Naturgesetz allenthalben gelte (GMS Ak. 04:423.33-37).

[11] (a) Denn ein Wille, der dieses beschlösse, (b) würde sich selbst widerstreiten, (c) indem der Fälle sich doch manche ereignen können, (d) wo er anderer Liebe und Teilnehmung bedarf, ... (GMS Ak. 04:423.37-38).

[12] I argue this thesis at greater length in "Kant's Self-Legislation Procedure Reconsidered" (2011; unpublished paper).

© Adrian Piper Research Archive Foundation Berlin

reformulate it as "a mere generality (*generalitas*) (2.15), in such a way that the practical principle of reason is supposed to join up halfway with the [derelict] maxim (2.16)."

Here Kant is targeting the polemical tactic that makes a virtue of inclination's sensible character, and a vice of reason's supersensible character; then embraces the former in order to reject the latter. To insist on the exceptional status of one's moral wrongdoing, based on the lack of empirical proof for the universality of reason's commands, is dishonest (KrV, A 747.28, A 748.25) and fraudulent (KrV, A 749.28). For there is in fact no empirical proof for that claim to exceptional status either, and the indexical strength of one's concrete, particular empirical inclination itself does not provide one. But of course the greatest victim of this deceit is oneself: Overcome by the empirical immediacy of inclination, and thereby persuaded that one has rationally decided to indulge it, one in fact abdicates one's rational autonomy, under the pretense that polemical reasoning provides. This, Kant feels, is to abandon one's dignity entirely (GMS Ak. 04:434.33, 440.02). This spectacle, of an agent wallowing in turpitude at the expense of reason, explains why moral wrongdoers often deserve not only our condemnation, but also our ridicule (think Charlie Chaplin on Hitler, or Will Ferrell on George W. Bush).

Kant's own critique of the Free Rider's faulty strategic reasoning in passage (2) does, indeed "dra[w] all of its decisions from the fundamental rules of [reason's] own formation (1.8)," just as passage (1) demands. It issues from the distanced perspective that is "oriented toward determining and judging the scope of entitlement of reason in general (1.4), according to the principles of its first institution (1.5)." The "principles of [reason's] first institution" provide Kant with the criteria of rule rationality relative to which any specific exercise of reason is to be critiqued. These criteria of rationality include the *universality* of these rules (KrV, A 88.09, 150.11-15), their *logical consistency* (KrV, A 150.11-15, A 151.01-08), their *objective necessity* (KrV, B 122.14; GMS Ak. 04:412.36), and their *conceptual unity* (KrV, A 103.01-110.09, B 359.01-10, A 650.25-29 and A 651.01-22, *passim*).

In passage (2), Kant does, indeed, apply precisely these criteria in critically analyzing the defensive self-justifications by which we try to protect our moral derelictions. Passage (2) invokes the *universality* of these rules in clauses (2.2), (2.4), (2.8), (2.9), and (2.14); their *logical consistency* in clauses (2.3), (2.7-10), and (2.12); their *objective necessity* in clauses (2.3) and (2.8); and their *conceptual unity* in clauses (2.5), (2.6), (2.10), and (2.12-15). These are the rational criteria relative to which Kant criticizes the Free Rider's reasoning as defective. That is, he brings the entire apparatus of the first *Critique* to bear on his analysis in the *Groundwork* of why free riding – i.e. any moral dereliction parasitic on others' rectitude – is irrational and self-defeating. He shows that it is inherent in the particularistic and self-exempting nature of purely self-interested reasoning to violate the objective requirements of reason. As he

© Adrian Piper Research Archive Foundation Berlin

earlier suggested (KrV, A 756.11-16), polemical self-justification is not properly a part of reason at all.

3. Kant' Second Solution: Promise-Keeping as a Perfect Duty

Passage (2) occurs as part of a more extended argument in Chapter II of the *Groundwork,* in which Kant attempts to make good on his supposition (GMS Ak. 04:421.10-12) that all imperatives of duty can be derived from the third (and fourth) formulation of the categorical imperative (GMS Ak. 04:421.07-09). He begins by introducing the four examples of moral dereliction with the comment that he is going to follow the traditional division into duties to self versus duties to others, and into perfect versus imperfect duties. In the footnote, he defines a perfect duty as one that "permits no exception to the advantage of inclination (GMS Ak. 04:421.27-28 and fn.). Then he discusses the four examples and the implications of trying unsuccessfully to universalize them (GMS Ak. 04:421.24-423.42). He comments that "[t]hese are some of the many actual duties, or at least some of those we take ourselves to have, whose derivation from the single principle mentioned above clearly come to mind (GMS Ak.04:423.43-44, 424.01-02)."[13] Kant claims, therefore, to have derived each one of the particular duties from the conceptual or volitional inconsistency produced by trying to universalize its transgression.

Next Kant generalizes these results, by drawing the distinction between a contradiction in the universalized *conception* of a derelict action, and a contradiction in the agent's *willing* of such a conception. This he equates with the distinction between strict, narrow or severe duty – that which he earlier called a perfect duty; and wider, meritorious duty – which he earlier called an imperfect duty (GMS Ak.04:424.02-17) respectively. Passage (2) then expands on Kant's concept of a contradiction in the will. In contrast to the purported clarity with which the derivation of these duties "come to mind," passage (2) illustrates the scattered mental gymnastics and subjective self-contradiction by which the Free Rider attempts to evade the force of this derivation, and excuse her self-exemption from it. Immediately following passage (2), Kant claims to have "determinately presented the content of the categorical imperative, which must contain the principle of all duty (if there is to be any),

[13] Dieses sind nun einige von den vielen wirklichen oder wenigstens von uns dafür gehaltenen Pflichten, deren Ableitung aus dem einigen angeführten Prinzip klar in die Augen fällt (GMS Ak.04:423.43-44, 424.01-02). The German Academy edition substitutes „Abtheilung" [division] for „Ableitung" [derivation]; and Timmerman's translation is true to this text whereas mine is not. The reason is the relation between this passage and GMS Ak. 04:421.10, where Kant uses and clearly means to use the verb „abgeleitet." If we take his pronouncement there at face value, then his use of „Abtheilung" at GMS Ak. 04:423.43 may have been a slip of the pen.

© Adrian Piper Research Archive Foundation Berlin

clearly and for every application (GMS Ak. 04:425.05-08)."[14] So in the context of Kant's derivation, passage (2) plays an essential role. It functions as the real-world counterpoint to the idealized theory; and details our willful but confused, *de facto* deviations from the ideal of deductive rationality that, on Kant's view, the categorical imperative itself expresses.

Many (myself included) have called into question the success of Kant's derivation of particular duties from the categorical "imperative."[15] But I have tried to show that the significance of passage (2) does not depend exclusively on its role in this attempted derivation. Rather, it also must be understood as a direct application of the argument of passage (1). There Kant claimed that a "critique of reason itself (1.19)" would resolve "a merely insecure peace granted by a mediating authority (1.14)" into "the calm of a lawful condition (1.10), in which we are to conduct our disputes in no other way than through the *legal process* (1.11)." I have tried to trace the ways in which passage (2) provides precisely such a critique of reason, namely of the Free Rider's reasoning. Kant exposes the polemic by which the Free Rider justifies his parasitic indulgence of personal inclination at the expense of the moral principles that others follow.

Now suppose Kant's critique does, in fact, have the claimed subduing effect on the Free Rider's internal vacillations between self-exemption from and fidelity to universal law. Suppose it therefore halts our internecine disputes over whose self-aggrandizing agenda is to take precedence, and how to exploit those on whose conformity to law our free riding is parasitic. Does this suffice, "because it affects the very source of the disputes themselves (1.17)," to deliver us permanently from the "state of nature … a state of injustice and violence (1.21)," into a condition in which we all voluntarily "submit ourselves to legal constraints (1.23) that limit our freedom (A1.24) so that it can be consistent with others' freedom (1.25) and the common good of all (1.26)"? Apparently not. Reminders of one's irrationality do not often count for much while one is in the grip of inclination; and reminders of one's bad faith do not often count for much while one is in the grip of self-interest. Free Riders do not mind looking ridiculous, as long as they get what they

[14] [I]mgleichen haben wir, welches schon viel ist, den Inhalt des kategorischen Imperativs, der das Prinzip aller Pflicht (wenn es überhaupt dergleichen gäbe) enthalten müßte, deutlich und zu jedem Gebrauche bestimmt dargestellt (GMS Ak. 04:425.05-08).

[15] *Op. cit.* Footnote 12. For some other recent treatments, see Freyenhagen, Fabian, "The Empty Formalism Objection Revisited: Par. 135R and Recent Kantian Responses," forthcoming in Brooks, Thom (Ed.), *Hegel's Philosophy of Right: Essays on Ethics, Politics, and Law* (Oxford: Blackwell, forthcoming 2011), and in revised form in the *Bulletin of the UK Hegel Society*; and Geiger, Ido, "What is the Use of the Universal Law Formula of the Categorical Imperative?" *British Journal for the History of Philosophy 18*, 2 (271-295).

© Adrian Piper Research Archive Foundation Berlin

want. These reminders merely drive the Free Rider deeper underground, to even more elaborate polemical subterfuge in the service of inclination. The more urgent challenge is how to loosen the grip of inclination and self-interest themselves.

Kant has some additional resources for addressing this challenge. The second example of moral transgression in the above derivation formulates the case very concretely, as one of intentionally breaking a promise to repay a loan. In the course of the derivation, Kant infers that this violates our perfect duty to others to always, without exception, repay our loans. However, he does not generalize from this description to any conclusions about the type of duty of which repaying a loan would be a token. Kant's earlier discussion of this case in Chapter I of the *Groundwork* (GMS Ak.04:402:19-403:21) reveals his ambivalence about how exactly to generalize and classify such a case. He describes it variously as "making a promise with the intention of not keeping it (GMS Ak. 04:402:20)," as a "false promise (GMS Ak. 04:402:23)," a "lie (GMS Ak. 04:402:27, 403:13)," a "lying promise (GMS Ak. 04:403:04; also see 429.35)," and an "untrue promise (GMS Ak. 04:403:10-11)." And his first formulation of the contradiction test in Chapter I refers to it as both a lie and the negation of a promise, which in turn produces a contradiction, both in its universalized conception, and also in the agent's will:

> (7) (1) Thus I soon become aware that I can indeed will to lie, (2) but not a universal law to lie; (3) for in accordance with such a law, (4) there actually would be no promises at all, (5) because it would be futile to profess my will in regard to future actions to others, (6) who would not believe this claim, (7) or, if they did over-hastily, would repay me in like coin; (8) consequently my maxim would destroy itself as soon as it were made a universal law (GMS Ak.04:403.12-21; cf. also 422.18-44).[16]

When he later reconsiders the second example in light of the seventh formulation of the categorical "imperative" (GMS Ak. 04:429.12-14), he refers to this "lying promise" as violating a necessary or culpable duty – presumably a perfect duty – to others (GMS Ak. 04:429.34-36).[17] It would seem

[16] So werde ich bald inne, daß ich zwar die Lüge, aber ein allgemeines Gesetz zu lügen gar nicht wollen könne; denn nach einem solchen würde es eigentlich gar kein Versprechen geben, weil es vergeblich wäre, meinen Willen in Ansehung meiner künftigen Handlungen anderen vorzugeben, die diesem Vorgeben doch nicht glauben oder, wenn sie es übereilterweise täten, mich doch mit gleicher Münze bezahlen würden; mithin meine Maxime, sobald sie zum allgemeinen Gesetze gemacht würde, sich selbst zerstören müsse (GMS Ak.04:403.12-21).

[17] Hobbes' original characterization of the Foole, or Free Rider, shows this same ambivalence about whether to conceptualize the case as one of breaking one's covenants (114), or deceiving others (115) for self-interested reasons. However, Hobbes treats the violation of covenant as itself a speech act that "consequently declareth that he thinks he may with reason do so …. (115)." *Op. cit.* Footnote 3.

© Adrian Piper Research Archive Foundation Berlin

that Kant is unsure whether to classify the violation of one's commitment to repay a loan as a species of lie, or as a species of broken promise; and whether to classify a lie as a species of broken promise, or a broken promise as a species of lie.

On the face of it, false promising would clearly seem to be that species of lie in which one deliberately misrepresents one's intention to act.[18] Furthermore, Kant's later claim that lying is morally unacceptable under any circumstances, even when the murderer is at the door,[19] shows that he regarded truth-telling as a perfect duty. Nonetheless, in this he was mistaken, even according to his own criteria. What transgresses our perfect duty to others in the *Groundwork* is not a lie, nor should it be. For I can both will to lie, and also will a universal law of lying, without engendering either a contradiction in my will, or a contradiction in my conception of this universalized law. I avoid a contradiction in my will, by willing that everyone else tell those same lies that echo and reinforce the lies I tell both myself and them. These may be as plentiful in number, scope and content as the extent of my own self-deception. Thus I can consistently will both to lie, and also a universal law of lying. I stand to lose nothing in a world in which my lies and everyone else's are mutually supporting. Similarly, I avoid a contradiction in my conception of that universalized maxim, by willing universalized lying that is as systematic, coherent, and mutually reinforcing as those lies I tell both myself and them. There is no necessary inconsistency in the conception of a world in which everyone lies, both to themselves and others, all the time. Nor would it be futile for any particular agent to do so, provided that everyone else did so as well. Of course such a world would be "detached from reality," so to speak. But that fact hardly bespeaks its conceptual impossibility.

What such a world cannot contain, on pain of conceptual inconsistency, is a universal practice of making false promises – that particular species of broken promise in which one knows at the time of verbally committing

---

[18] I confine this discussion to the core definition of a *lie* as an assertion that the speaker recognizes to be false, made with the intention to deceive the listener. Thus I leave aside intentionally misleading but true assertions that deceive through circumlocution, non-assertional actions or gestures that deliberately deceive through suggestion, intentionally withheld true assertions that deceive by encouraging false inferences, etc.; as well as of how recognizing an assertion to be false is related to knowing, believing, perceiving or conceiving it to be false. On this last point, see my *Rationality and the Structure of the Self, Volume II: A Kantian Conception*, Chapter VII. "Pseudorationality," *op. cit.* Footnote 4.

[19] Kant, Immanuel, "Über ein vermeintes Recht aus Menschenliebe zu lügen," *Kant's Gesammelte Schriften, Band VIII: Kant's Werke: Abhandlungen nach 1781*, Herausg. Königlich Preußischen Akademie der Wissenschaft (Berlin und Leipzig: Walter de Gruyter & Co., 1923), Ak. 08:423-430.

© Adrian Piper Research Archive Foundation Berlin

oneself to the action that one will be unable to fulfill that commitment. The contradiction in conception argument Kant offers in passage (7) applies correctly here; and does, indeed, rule out false promising as a universal practice. One makes the false promise in order to raise expectations of its fulfillment in the other, and in order to cause the other to perform actions based on those expectations that are beneficial to oneself. But one loses all the advantages of false promising if everyone does it. For in that case, the utterance of a promise has no performative force; hence its falsity is easily detected. Whereas skillful lying need never be exposed by its falsifying conditions, false promising is regularly exposed by its falsifying omissions. Universal false promising then would be a settled linguistic practice to which no promised performance ever corresponded. It would function as a series of meaningless utterances that failed to raise the requisite expectations in others, and hence would fail to count as promising at all. So if false promising is a species of lying, it lacks an important property that the higher-order concept of lying has, namely universalizability. In this case, the Free Rider induces in herself not merely a contradiction in her will, but a contradiction in her conception of what it is she is intending to do. She thereby short-circuits the very powers of reasoning she meant to harness in the service of self-interest.

The conceptual impossibility of universalizing a practice of false promising is equivalent to the conceptual impossibility of a functioning society in which no one's word on any topic can be trusted. The source of this universal mistrust is not that everyone knows everyone else to be lying. It is entirely possible to trust others whom one knows are lying, by trusting them to lie and taking appropriate precautionary measures. In some cases, this may mean simply use *modus tollens* to infer the truth. In other cases, the liar's assertion may habitually trigger an independent fact-check. In yet others, extended informal exposure may reveal a systematic but non-denotative relationship between the liar's false assertions and the truths he conceals. Combine all of these variants and a rather familiar psychological stance – of wary skepticism – emerges. But none of them, either singly or in combination, necessarily engenders universal and thoroughgoing mistrust, because none of them necessarily disturbs settled conventions of linguistic reference.

The universal mistrust engendered by universal false promising arises, rather, because one's stated intentions raise no expectation in others that one will realize those intentions in action. Hence one's intentional use of language raises no expectation in others that one will realize those intentions in the speech acts one in fact performs. Therefore others find no reliable connection between what one intends and what one says, or between what one says and what one does. Universal false promising is much more socially destructive and conceptually unthinkable than mere lying. For unlike lying, universal false promising destroys any *systematic* connection between meaning and

© Adrian Piper Research Archive Foundation Berlin

linguistic practice. This makes it impossible for linguistic communication – including individual instances of false promising – to take place.

It is important to be clear about what is at issue here. A community may lack the necessary conditions for the veridical use of language, in case everyone systematically lies. But this does not entail the impossibility of successful communication. Utterances such as, "How lovely to see you!" or "I have decided to resign in order to spend more time with my family," or "The dog ate my homework," still have meaning. They just do not have the meaning these sentences prima facie express. A community lacks the necessary conditions for the use of language *überhaupt*, hence those for successful communication, in case no such utterances have any determinable systematic meaning at all; that is, in case there is no shared and reliable agreement on how these utterances are to function. Even lying presupposes such agreement – and therefore that the parties to the agreement abide by the linguistic conventions agreed. Hence lying, like linguistic communication more generally, is a species of promise-*keeping*, in which we keep an agreement to deploy linguistic conventions that systematically deviate from truth-telling. So it is not universal lying, but rather universal false promising that irreparably destroys the Social Contract.

Under conditions of universal false promising, there is thus no incentive to communicate at all – nor, it would seem, is there any way in which communication can be established.[20] One must simply observe others' behavior, including their utterances; make inductive generalizations as to its regularities; hypothesize its motives; and draw one's own conclusions as to how best to exploit that behavior in order to serve one's own interests. Because no such utterance can be trusted, no interpersonal connection based on it can be established. Hence no relationship based on that connection can be developed. The conditions necessary for human cooperation are absent. At best, others are perceived as useful pawns; at worst, as opaque and unpredictable enemies.

This is the condition indicated in passage (1) by Kant's remark that without rational critique, "reason is, as it were, in the state of nature (1.6), and can only validate and secure its claims and demands through *war* (1.7)." Since communication is impossible in the state of nature, any resort to techniques of

[20] There is a very large literature on this topic. For early discussions, see Hodgson, D. H., *Consequences of Utilitarianism: A Study in Normative Ethics and Legal Theory* (Oxford: Clarendon Press, 1967); Lewis, David, *Convention: A Philosophical Study* (Cambridge, Mass.: Harvard University Press, 1969); Gibbard, Allan, "Utilitarianisms and Coordinations" (Ph.D. diss., Harvard University, 1971); Ullman-Margalit, Edna, *The Emergence of Norms* (Oxford: Clarendon Press, 1977); Piper, Adrian M. S., "Utility, Publicity and Manipulation," *Ethics 88*, 3 (April 1978), 189-206; Regan, Donald, *Utilitarianism and Cooperation* (Oxford: Clarendon. Press, 1980).

© Adrian Piper Research Archive Foundation Berlin

rational argument is futile. Of course reason as an innate human faculty can operate within each agent. But its medium of outward expression must bypass language. Instead, the claims and demands of reason issue in direct action and direct aggression against resistant others, and therefore can be secured only through force. Kant seconds Hobbes' own claim about the quality and character of the state of nature:

> (8) [T]he nature of war, consisteth not in actual fighting; but in the known disposition thereto, during all the time there is no assurance to the contrary. …
>
> Whatsoever therefore is consequent to a time of war, where every man is enemy to every man; *the same is consequent to the time, wherein men live without other security, than what their own strength, and their own invention shall furnish them withal.* In such condition, there is no place for industry; … and consequently … *no arts; no letters; no society;* and which is worst of all, continual fear and danger of violent death; and the life of man, solitary, poor, nasty, brutish, and short [italics added]."[21]

Kant's argument against false promising in passage (7) depicted a condition of generalized enmity in which communication and cooperation fail (7.5), mistrust prevails (7.6) and mutual retaliation for betrayal is pervasive (7.7). This is, in effect, the same condition to which Hobbes' more forceful and vivid depiction in passage (8) refers. Kant's analysis of the failure of universalization formally models Hobbes' analysis of the failure of the Social Contract. [22] Were the Free Rider's attempted self-exemption universally adopted in the state of nature as Hobbes envisions it, no Social Contract could be established because, as Kant observes, "actually there then would be no promising at all (7.4)," and hence no shared rules to obey. This is the degraded condition that Kant's universalization requirement excludes. Failed states, polluted by corruption at every level of government and society, unable to secure even the most basic agreements among warring political factions or enforce even the most basic legislation, populated by roving Free Riders whose capacity to reason itself has been dimmed by the pull of

[21] *Op. cit.* Footnote 3, 100.

[22] This may explain Kant's motivation for the fourth, "law of nature" formulation directly preceding his discussion of the four examples (GMS Ak. 04:421.21-23). Kant remarks that "the universality of the law in accordance with which effects occur constitutes that which is actually called *nature* in the most general sense (regarding its form) … (GMS Ak. 04:421.16-18)". In the German original, "Weil die Allgemeinheit des Gesetzes, wonach Wirkungen geschehen, dasjenige ausmacht, was eigentlich *Natur* im allgemeinsten Verstande (der Form nach), ... heißt, usw." The point would be to envision universalizable principles of rational action as deterministic *laws* of nature that correct the disruptive and irrational "effects [that] occur" in the *state* of nature.

© Adrian Piper Research Archive Foundation Berlin

immediate self-interest or extinguished by the exigencies of perpetual crisis, illustrate concretely what such a state of nature would look like.

4. The Deductive Relationship between Kant's Two Solutions

Thus Kant refutes the Foole's reasoning even in improved strategic form, and also provides a concrete illustration of how his critique of pure reason might resolve interpersonal conflict in the political and moral arena. Passage (2) critiques the Free Rider's reasoning by showing that, quite independent of the actual, external social consequences on which Hobbes' analysis relied, false promising is *individually* irrational in this double sense: It fails to advance her self-interest, because it actively undermines the cognitive coherence of the self whose interest it is. This, Kant shows, is a direct consequence of the conflict between her attempts to particularize her situation on the one hand, and to universalize its principles on the other – a futile undertaking nevertheless necessitated by the cognitive functioning of human reason itself. This incoherence leaves the Free Rider with delinquent inclinations but no basis on which to exempt them from reason's requirements.

I have just argued in Section 3 that passage (7) (plus its later elaboration at Ak. 422) correctly concludes that promise-keeping is a perfect duty. So Kant's attempt to derive particular duties from the universalization requirement is in this case successful: "from this one [categorical] imperative" at least *one* "imperativ[e] of duty can be derived (GMS Ak. 04:421.9-10)." We have also seen in Section 3 that Passage (2) is that premise in the derivation that critiques our real-world deviation from the ideal of deductive rationality into pseudorational polemic and incoherent self-justification, from the distanced standpoint of that ideal. However, we saw in Section 2 that passage (2) itself applies a premise not found in the *Groundwork*, but rather imported from passage (1) in the first *Critique*. The basic premises of Kant's derivation of promise-keeping as a perfect duty thus begin not with his brief summary of universalization in the *Groundwork*; but rather with the full-blown criteria of rationality, and the need to evaluate individual deliberation in their terms, that he spells out in the first *Critique*. The relationship between Kant's first, first *Critique* solution – the critique of reasoning, and the second, *Groundwork* solution – always to honor one's promises, is therefore one of *ceteris paribus* implication.

I have already tried to indicate the value of the first solution. But I have also argued that rational critique faces obstacles in its battle against empirical inclination. Does the second have a role or function independent of its status as derivative from the first? For example, might Kant's injunction in passage (7), that promise-keeping is a strict, necessary duty that "permits no exception to the advantage of inclination," have a pragmatic role in reinforcing the impact of the first on delinquent inclinational tendencies to incoherent self-

© Adrian Piper Research Archive Foundation Berlin

justification? Or might this injunction perhaps function to strengthen our disposition to "submit ourselves to legal constraints (1.23) that limit our freedom (1.24) so that it can be consistent with others' freedom (1.25) and the common good of all (1.26)"? And are we, in fact, moved by the counterfactual spectre of betraying our word of honor to critically monitor the warped deliberations meant to excuse those betrayals?

Consider two rather different ways of thinking about Kant's injunction. First the cheerful variant: One effective strategy for restoring trust betrayed is to earn it; to demonstrate, through one's own actions, that one can, in fact, be trusted. Promise-keeping is a ready tool, though not the only one, for demonstrating this. When I make a promise, I voluntarily place myself under an obligation to perform some future act. When I keep a promise made silently to myself (no more chocolate-covered cherries, I tell myself sternly), I demonstrate my trustworthiness – or lack thereof – to myself alone. When I keep a promise made to myself publicly (by donning a nun's habit, say), I demonstrate my degree of trustworthiness and resolve before the gaze of disinterested spectators who may or may not cheer me on. When I make the promise to another (by signing a contract, for example), I call forth his interest and psychological investment in my ability to demonstrate trustworthiness and resolve to him. But in all of these cases, the only person ***I*** need to trust is myself.

Trusting oneself is far from self-evident, as Descartes instructs us; and the mixture of hope, doubt, uncertainty and self-deception about one's ability to deliver that comes with doing nothing is a soporific. In order to be fit to earn others' trust, I must first earn my own, by keeping the promises I make to myself. A solid track record here gradually builds the self-confidence in my own trustworthiness and resolve that I need in order to survive others' skeptical scrutiny. A failure of will at this point is, literally, demoralizing; and can be fatal. For a failure to pass my own trust test manifests an inability of my present directive to govern my subsequent behavior, and so a disjunction between speech and action that, if too often replicated, may shade off into schizoid dissociation. I can minimally trust myself only if my behavior accords with my assertions. To keep a promise is to do what I have said I am going to do, because I have said I am going to do it. If I cannot trust myself to do that much, I cannot expect trust from anyone else.

The more promises I am able to make and keep, the more trust I can, indeed, expect from others. The more secure are the social connections engendered from them, the more durable the social fabric they weave. This can be a particularly useful practice when the Social Contract has been so badly damaged that we are, in fact, reduced to merely observing others' behavior, disregarding the face value of their verbal pronouncements, and drawing our own conclusions privately. For all promise-keeping requires of each promisor is that one be willing to submit one's own behavior to this very

© Adrian Piper Research Archive Foundation Berlin

same scrutiny, as soon as one has built enough self-confidence to know one can pass its test. At that point, Kant's command that we are to keep *all* of our promises – and hence to make none we cannot keep – may not seem so unrealistic. It may even have its own peculiar charm.

However, the cheerful variant on Kant's command is not a panacea for repairing the Social Contract, for it rightly and predictably earns the *mis*trust of those who subordinate it to unconditional personal or professional loyalty. One who demands your support or obedience even for unconscionable behavior will justifiably regard your insistence on keeping your promises regardless as betrayal or insubordination, or in any case as threatening her interests; and so will justifiably regard you as unreliable and untrustworthy precisely because of your fidelity to the moral law. This is the source of the enmity and vengefulness with which whistle-blowers are treated. Where free riding is rampant, the whistle-blower betrays the trust of those who benefit from it, by insisting on the priority of moral injunctions that – as we have seen in Section 2 – the Free Rider makes every effort to disregard. So there are many circumstances under which cultivating a disposition to promise-keeping will exacerbate mistrust – at least of oneself by others – rather than heal it. Thus far, it would seem, Kant's second solution, at least under the cheerful interpretation, provides no Foole-proof instrumental corrective to incipient social disorder at all. Quite the opposite.

So we need to adjust our conception of our own condition downward in order to accommodate this fact; and consider a second interpretation of Kant's command. The depressing variant runs as follows. The point has been made very often that once entrapped in a state of nature, it is virtually impossible to extricate oneself from it.[23] Section 2 above contended that our actual, immediate challenge is not to extricate ourselves from the state of nature, but rather to prevent our descent into it – that is, to prevent the proliferation of free riding to such an extent as to actually constitute a universal or near-universal practice. Each of us risks such a descent, in so far as the personal advantages we reap from our relations with others are parasitic on our betrayals of their trust. When we debase the words of honor we regularly offer them – to keep a confidence, to speak forthrightly, to stand by them in adversity, to abide by our agreements, each of these individual betrayals corrode our personal, familial, and social relationships, as well those we form in the workplace, in civic involvement, in politics and in law. Each time we betray others' trust, we teach them that we cannot be trusted and that words cannot be trusted. Each time we are on the receiving end of such betrayals, we

[23] See any discussion of the Prisoner's Dilemma. Steven Kuhn's article, "Prisoner's Dilemma," Stanford Encyclopedia of Philosophy (2007; http://plato.stanford.edu/entries/prisoner-dilemma/) contains a good discussion and bibliography. Also see Note 18, above.

© Adrian Piper Research Archive Foundation Berlin

learn these lessons ourselves. The more we learn, the more our behavior adapts accordingly; the more fully we habituate ourselves to this self-inflicted condition; and the more we drag ourselves and our rational capacities off course by excusing or pseudorationalizing the descent into mutual mistrust that both Hobbes and Kant so eloquently deplore.

The more we indulge such derelict intentions, the more we lose our grasp on the cognitive and social danger they – and, increasingly, we – represent. In this way, we asymptotically approach that cognitively degraded condition experienced by the citizens of a failed state, in which the exigencies of immediate survival dull one's powers of imagination and ratiocination, and so one's insight into the implications and long-term consequences of one's own behavior. From this dimmed and narrowed perspective, Kant's injunction to always keep our promises – and therefore, of course, to make none we anticipate being unable to keep – is scarcely thinkable, much less credible. For the ability to understand what a promise is, what it entails, and why it must be honored requires this very ability to anticipate in the first person case – to predict and infer, hence to theorize and universalize over the first personal particular – that is being eroded through disuse.

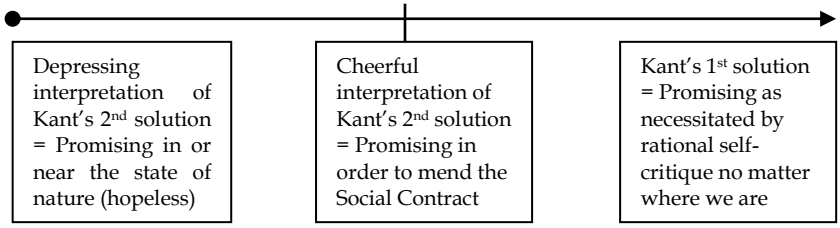
This is precisely the failure of vision described in Section 1, of which Kant implicitly accuses the Free Rider in passage (2). The Free Rider does not necessarily lack the capacity to criticize *other* people's behavior or reasoning with reference to consistent and universal moral principles. The Free Rider's deficit consists in an inability or unwillingness to apply those same principles to a conceptualization and evaluation of her *own* behavior and reasoning. Her failure of vision is a failure to jettison the standpoint of her own inclinations in favor of the standpoint of her own critical rational faculties. It is a failure of self-reflection and self-criticism, without which it is not possible to grasp why keeping one's own promises might be an important or valuable thing to do.

So these two interpretations of Kant's second solution as independent of its derivation from the first would seem to stand at opposite ends of a continuum. At the cheerful end, Kant's injunction to the perfect duty of promise-keeping is an invitation to cultivate a practice of earning trust as a means of mending the Social Contract. At the depressing end, it is a feeble thought-experiment so chimerical and remote from our actual social condition as to be not merely unrealistic but practically unthinkable. This is the case in which Kant's injunction seems either meaningless or ridiculous or pragmatically unacceptable. Whether or not we are far enough along the cheery stretch of the continuum to take the risk of rejecting the depressing interpretation out of hand, as surely inapplicable to our surely not so entirely hopeless actual circumstances, is a moot question.

But even to cultivate the practice of promise-keeping as an instrument for earning trust and thus mending the Social Contract is not possible unless one has already grasped the conceptual significance of this practice as a good in

© Adrian Piper Research Archive Foundation Berlin

itself. For the practice presupposes the discipline of self-interrogation and self-criticism that Kant's first solution supplies. So the cheerful interpretation of Kant's second solution in fact does not stand at one end of the continuum, but rather midway between the depressing interpretation at one end, and Kant's first solution at the other:



In order for Kant's second solution to reinforce the impact of the first, it must already have the support of the first and follow from the first. It is futile to command anyone, including oneself, to keep one's promises, if one lacks the capacity to grasp the import of doing so. Under these circumstances the project of earning trust cannot even get off the ground. The capacity one needs to grasp the significance of honoring one's promises is precisely the capacity to reason critically and reflectively, with attention to the disparities between our spontaneous deliberations and the criteria of rationality that they aspire to fulfill. Unless we can compare our pseudorational self-justifications for breaking our promises with the rational principles they violate, and condemn those violations from their distanced perspective, we cannot recognize our moral failures, and therefore cannot recognize the social disorder that each one of these betrayals singly demonstrates to others. Hence we cannot appreciate the significance of Kant's injunction to avoid such betrayals at all costs. Delinquent inclinations that are unresponsive to Kant's first solution are, by definition and in fact, beyond the reach of the second.

So Kant's second solution to the Free Rider problem, that promise-keeping is a perfect duty, is, after all, a straightforward implication of the first, and not an independent support for it. Supplementing the derivation in the *Groundwork* with this additional premise from the first *Critique* increases its plausibility. We need to see, not merely that we must not betray others' trust; and not even merely that we could not want or conceive a world in which everyone did so. We also need to see that we *dare* not betray our*selves* with trumped-up excuses or self-aggrandizing pseudorational gymnastics about why it is acceptable to do so despite this. Once we have freed ourselves from the Free Rider's cramped vision sufficiently to acknowledge all of these things, we are in a position to appreciate the significance of Kant's injunction that we are never to dishonor our promises. Only then can we recognize the whistle-blower's "betrayals" for what they are.

© Adrian Piper Research Archive Foundation Berlin

To claim that Kant's second solution depends on the first, rather than the other way around, is also to claim that the effectiveness of the second depends on individual circumstance in a way that the first does not. We all have powers of deliberation that would benefit from self-conscious rational critique, and we all know or can discern the criteria relative to which to critique them. But we do not all have the blessings of fortune or circumstance necessary to exercise those powers fully or well – nor, therefore, the wherewithal to grasp the significance of promise-keeping or to enact it in practice. Even the self-confidence needed to make a promise to oneself, silently, in the justified expectation that one will be able to keep it, is a resource in very short supply; and often a casualty of the same institutional or familial betrayals of trust it should be recruited to heal. The demonstrations of trustworthiness needed to earn trust themselves presuppose a Social Contract at least intact enough to support them. Others must be at least curious enough about whether or not we can be trusted to observe our attempts to demonstrate that we can. They cannot be so damaged by their own experiences of betrayal as to write us off, cynically, before we have even had a chance to try to earn their trust; or to write themselves off, despairingly, before giving themselves a chance to earn ours.

Each of us must estimate for ourselves the proportion of our daily lives that is dominated and shaped by the betrayals of trust that false promising effects, and hope we estimate correctly. But each one of us who is likely to read this essay almost certainly has, in fact, had direct and irrefutable experience of a society in which the Social Contract has been very badly damaged indeed – not only by politicians but also by friends, family, colleagues and institutional representatives of every stripe, as well as by oneself. Correspondingly, each one of us must find a way to assess the damage this has wrought, in turn, on our ability and disposition to repair it. When we permit ourselves to survey the vastness and complexity of this damage, to ourselves as well as to the Social Contract, we may feel despair at the inadequate resources – of time, energy, imagination, motivation – any one of us can individually contribute to the project of restoring a shared foundation of mutual trust to which our own experiences may be inadequate. Whether the degree of despair we feel is in fact congruent with our ability to repair the Social Contract is a matter each of us must settle for ourselves. Yet repairing that damage may not be as daunting as it seems, nor the individual task for each one of us so overwhelming. After all, if breaking our promises has the foundationally destructive social role Kant has claimed, then keeping them, one promise at a time, would at least reverse the trend.

© Adrian Piper Research Archive Foundation Berlin

References

Freyenhagen, Fabian, "The Empty Formalism Objection Revisited: Par. 135R and Recent Kantian Responses," forthcoming in Brooks, Thom (Ed.), *Hegel's Philosophy of Right: Essays on Ethics, Politics, and Law* (Oxford: Blackwell, forthcoming 2011), and in revised form in the *Bulletin of the UK Hegel Society*

Geiger, Ido, "What is the Use of the Universal Law Formula of the Categorical Imperative?" *British Journal for the History of Philosophy 18*, 2 (271-295)

Gibbard, Allan, "Utilitarianisms and Coordinations" (Ph.D. diss., Harvard University, 1971)

Hobbes, Thomas, *Leviathan*, Ed. Michael Oakeshott (New York: Macmillan/Collier Books, 1977)

Hodgson, D. H., *Consequences of Utilitarianism: A Study in Normative Ethics and Legal Theory* (Oxford: Clarendon Press, 1967)

Kant, Immanuel, *Kritik der reinen Vernunft*, Herausg. Raymund Schmidt (Hamburg: Felix Meiner Verlag, 1976)

_____, *Grundlegung zur Metaphysik der Sitten*, Herausg. Karl Vorländer (Hamburg: Felix Meiner Verlag, 1965)

_____, *Groundwork of the Metaphysic of Morals: A German-English Edition*, trans. and ed. Mary Gregor and Jens Timmermann (Cambridge, UK: Cambridge University Press, 2011)

_____, *Groundwork of the Metaphysic of Morals*, trans. H. J. Paton (New York: Harper Torchbooks, 1964)

_____, "Über ein vermeintes Recht aus Menschenliebe zu lügen," *Kant's Gesammelte Schriften, Band VIII: Kant's Werke: Abhandlungen nach 1781*, Herausg. Königlich Preußischen Akademie der Wissenschaft (Berlin und Leipzig: Walter de Gruyter & Co., 1923), Ak. 08:423-430

Kuhn, Steven, "Prisoner's Dilemma," Stanford Encyclopedia of Philosophy (2007; http://plato.stanford.edu/entries/prisoner-dilemma/)

Lewis, David, *Convention: A Philosophical Study* (Cambridge, Mass.: Harvard University Press, 1969)

© Adrian Piper Research Archive Foundation Berlin

Piper, Adrian M. S., "Utility, Publicity and Manipulation," *Ethics 88*, 3 (April 1978), 189-206

_____, *Rationality and the Structure of the Self, Volume II: A Kantian Conception*
(http://www.adrianpiper.com/rss/docs/Rationality%20and%20the%20Structure%20of%20the%20Self,%20Volume%20II-%20A%20Kantian%20Conception.pdf , 2008)

_____, "Kant's Self-Legislation Procedure Reconsidered (unpublished paper, 2011).

Regan, Donald, *Utilitarianism and Cooperation* (Oxford: Clarendon. Press, 1980)

Sen, Amartya K., "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory," *Philosophy and Public Affairs 6*, 4 (1977), 317-44.

Ullman-Margalit, Edna, *The Emergence of Norms* (Oxford: Clarendon Press, 1977)

© Adrian Piper Research Archive Foundation Berlin