

## UNKNOWNABLE OBLIGATIONS\*

You face two buttons. Pushing one will destroy Greensboro. Pushing the other will save it. (Sorry, doing nothing also destroys Greensboro.) There is no way for you to know which button saves and which destroys. What ought you to do? Answer: You ought to make the correct guess and push the button that saves Greensboro. Second question: Do you have an obligation to push the correct button?

Most people answer no on the grounds that all obligations are weakly self-intimating:

(Access) If one is obliged to do x, then one can know one is obliged to do x.

My negative goal is to refute access. Since 'can know' can be relativized to different background constraints, I advance counterexamples that range from practical impossibilities to logical impossibilities. As with other critiques of self-

intimation principles, I will be reacting to the readings set by defenders.

As the paper progresses, I will be increasingly guided by readings suggested by the positive theme that many obligations must be satisfied blindly. Our degree of access rises continuously from this zero point in a pattern similar to one psychologists have recently advanced for perception. Much moral knowledge is a "bag of tricks": a motley of low level, low reliability, domain specific, heuristics that satisfies the economics of information by their collective reliability rather than their individual performance.

This naturalistic picture of moral epistemology will be further motivated by an internal criticism of inwardly oriented ethical systems, such as Immanuel Kant's. By asserting the primacy of a good will, these systems become embroiled in the incoherency that drives Gregory Kavka's toxin puzzle. The puzzle also brings out the role of limits of belief and intention in deterrence. Retributivists have mischaracterized these limits as deontic in their classic objection about punishing the innocent. Thus action theory narrows the range of feasible theories of obligation while undercutting an influential

objection to one of the members of this range, namely, objective act utilitarianism.

## I. THREE ARGUMENTS FOR SELF-INTIMATING OBLIGATIONS

Most people find access plausible on its own. If pressed, they adduce three arguments:

### A. Unfair Punishment

People should have an opportunity to reliably comply with ethical demands. It would be Kafkaesque injustice to penalize someone for violating a undetectable prohibition. Morality, at least, must be fair. Besides morality guides action.

Unknowable directives undermine the whole point of shepherding people into good behavior. A charitable interpretation of a moral code should not saddle it with obvious irrationalities -- such as commanding without communicating those commands.

### B. 'Ought' implies 'Can'

A related defence of access is that it is entailed by the principle that 'ought' implies 'can'. Obligations entail 'ought'

statements. So if I cannot know where duty lies, how can I do what I ought?

Well, by luck. Don't syllogize "Knowledge is power, I have no knowledge, therefore, I have no power over Greensboro's fate". "Knowledge is power" uses the 'is' of predication, not identity. Some power is not knowledge. After all, in the button case you have a 50/50 chance of saving Greensboro. If you were tied to your chair, then you could not push the right button. But your hands are free.

The 'ought' implies 'can' objection can be revived by appealing to a thicker reading of the principle:

(R) If one ought to do x, then one can reliably do x.

Unlike the thin reading, (R) says that the performance of one's obligations must be within one's control. And indeed, if knowledge of one's obligations is the only means of non-accidentally satisfying one's obligations, then acceptance of (R) will lead to acceptance of access.

\

### **C. The Publicity Principle**

The second appendix of Immanuel Kant's "Perpetual Peace" presents a "transcendental formula of public law": "All actions relating to the rights of other men are unjust if their maxim is not consistent with publicity." This requirement of openness flows smoothly from the categorical imperative insofar as it requires us to act in accordance with just those laws that we would be willing to enact as a rational being for a kingdom of ends. It is more than a ban on secret policies such as 'Negotiate with terrorists only when this willingness can be concealed'. Knowledge should enhance what is known by providing a framework of mutual expectations. Ideally, everyone should know the laws, know that everyone knows the laws, and so on. It is this element of common knowledge that enables ethics to blossom into political theory:

The parties assume that they are choosing principles for a public conception of justice. They suppose that everyone will know about these principles all that he would know if their acceptance were the result of an agreement. Thus the general awareness of their universal acceptance should have

desirable effects and support the stability of social cooperation. (Rawls 1971, 133)

The proponent of access can plausibly claim support from the publicity principle. After all, if obligations are made binding by an agreement or an inner act of universal affirmation, then these obligations must be known by the agents. An approver must be aware of what he is approving.

However, the publicity principle is typically intended as a constraint on a small set of foundational principles governing an idealized set of agents in a hypothetical situation. Empirical uncertainties appear to block any inference from publicity to access. For example, the difference principle only states that social and economic inequalities must be arranged so that they are to the greatest benefit to the least advantaged (Rawls 1971, 302). Insurmountable problems about measuring social and economic status might make a mystery of what counts as satisfying this obligation.

A proponent of access would recoil at the suggestion that we often can do no more than guess at where duty lies. The difference between those who satisfy their obligation and those who violate their obligation would then be a matter of

luck (Williams 1976). A good conscience would not make one safe from censure.

Moreover, Kant regards the publicity condition as a practical guide for concrete actions rather than a mere theoretical constraint for screening principles. Think of how existentialists echo this theme of self-legislation: when a man chooses, he chooses for mankind. Had any other person been in my position (the situation as I perceive it), then I would have been willing that they make the same choice.

One can preserve this concreteness by subjectivizing the clauses of the social contract. Instead of being obliged to favor the least advantaged, one's obligation is limited to doing what one perceives to be favoring the least advantaged. In Kantian terminology, social contract principles are regulative ideals rather than proper obligations. Our real obligation is to aim at the satisfaction of these ideals. The actual achievement of the goal is not required. Indeed some regulative ideals are impossible to satisfy. Requiring their satisfaction rather than their pursuit would violate Kant's dictum that 'ought' implies 'can'.

The proponent of access could acknowledge an epistemological asymmetry between abstract principles and

concrete obligations. Abstract principles are hypothetical common knowledge in the sense that any rational agent could deduce them a priori. However, common knowledge of concrete obligations is doubly hypothetical. For concrete obligations need a posteriori background. If that background is assumed and if one does the reasoning, then one will arrive at same conclusion. Hence, anyone adopting the agent's perspective should arrive at the conclusions. Thus even one's particular obligations must withstand meet a publicity condition: had your particular circumstances been known, others would have been able to know what your obligation was. There is no private moral knowledge.

Given a strong but plausible modal logic, such as S4, what is possibly possible is possible. Hence the doubly hypothetical nature of a publicity condition for concrete obligations would still leave it in conformance with a principle even stronger than access: any obligation can be known by anyone. So I think that the proponent of the access would be correct in expecting support from the publicity principle. Therefore, my opposition to access spreads to the publicity principle.

Objective utilitarians are renowned for their violation of the publicity principle. They acknowledge that there are circumstances in which the greater good is served by people not believing in utilitarianism. Henry Sidgwick (1907, Book 4, Chapter 5, Sections 2 and 3) suggested that ordinary people might find utilitarianism so abstract that they would not be motivated to comply. Utilitarian leaders might therefore encourage belief in codes that are more compelling than utilitarianism.

This intellectual paternalism has been condemned as intolerably manipulative (Piper 1978). Kurt Baier has even criticized esoteric morality as contradictory; moral principles are "meant to be taught to all members of the group in such a way that everyone can and ought always to act in accordance with these rules" (1958, 196). Bernard Williams (1973, 135) has attacked "government house utilitarianism" as at least pragmatically self-defeating. If utilitarianism implies that no one ought to believe it, then it surrenders its founding ambition of reform through reason and calculation.

In addition to these explicit defenses of the publicity principle, we should add the voices of those who tacitly rely on it. Karl Marx, Frederich Nietzsche, and Sigmund Freud

attacked the moral codes of the Bourgeois, Christianity and the Victorians by showing how these codes owe their existence to ignorance or irrationality (Slote 1977). These unmaskers are preoccupied with the exposure of myth, "false consciousness", "repression", etc.

## II. ACCESSIVE ABSURDITIES

Despite supporting arguments and intuitive plausibility, the access principle has awkward consequences. At a logical level, the access principle blesses one species of argumentum ad ignorantiam:

It is impossible for me to know that I am obliged to do x.

Therefore, it is impossible for me to be obliged to do x.

This argument might seem acceptable to those who sharply distinguish between normative and descriptive knowledge.

They think that the epistemology of morality, language, and logic is dramatically different from the epistemology of empirical knowledge. Although aspects of nature may be hidden by their size or distance or complexity, it seems that

normative matters must be accessible because we are the creators and targets of those norms.

### **A. The Publicist meets the Holist**

This epistemological echo of the fact/value distinction is untenable. Only mild epistemological holism is needed to link a normative publicity principle to a descriptive publicity principle. For if my obligations must be knowable, any necessary background facts must also be knowable.

This spill-over can be illustrated by the atheist's appeal to intellectual evil. God's existence would be a crucial background fact for anyone trying to ascertain their obligations. So by the access principle, God's existence would be knowable to every human being. But at least some people cannot know that God exists. (For a similar argument, see Schellenberg 1993.)

Some theists believe that everyone can indeed know that God exists. Anselm noted that anyone who has wit enough to deny God's existence must have the idea of God. And according to Anselm, the idea of God alone is enough to prove God's existence via the ontological argument. Rene Descartes thought that the presence of the idea of God could only be

explained as an effect of God. Knowledge of God's omnibenevolence plays a pivotal role in Descartes' proof of an external world: our ideas of the world must be informative enough to ensure that we have a fair opportunity to learn about the world. Otherwise God would be a deceiver.

Regardless of whether one finds the above arguments cogent, they are counterexamples to the comforting assumption that the access principle can be confined to the purely normative realm.

### **B. Self-intimation precludes indeterminacy**

The access principle would make certain kinds of dilemmas too easy to resolve. Suppose Larry knows the access principle and also knows that it is good to save the life of a stranger by mouth to mouth resuscitation. However, he is not sure whether such aid is obligatory or supererogatory. After much research he concludes that he cannot know that the aid is obligatory. Larry then conjoins this lemma with the access principle to deduce that the aid must therefore be supererogatory. Hasty inference!

Anti-dilemma lemmas also invade the arena of moral conflicts. Suppose that Sophie knows that either she is obliged

save her son or that she is obliged not to save her son (because her son can only be saved by sacrificing her daughter). Sophie also knows she cannot know which obligation holds because there is no moral difference between her son and her daughter. A contradiction follows if Sophie can know the access principle. For she can then reason ad ignorantiam that she is not obliged to save her son and then reason ad ignorantiam that she is not obliged not to save her son. This contradicts the original supposition that she knew that one of these obligations hold. Thus the access principle precludes alternatives that are borderline between being obliged and forbidden. However, the most poignant cases in the literature on moral dilemmas have precisely this character. (In an earlier paper (Sorensen 1991) I argued that acts which appear to be both obligatory and forbidden are just cases that are borderline between 'obligatory' and 'forbidden'. Since I also espouse an epistemic theory of vagueness, this committed me to viewing many apparent moral dilemmas as only involving unknowable obligations.)

There is a parallel with Roderick Chisholm's (1942) speckled hen objection. Phenomenalists believed that sense data were (at least weakly) self-intimating -- if one's sense

datum had a certain feature then one could know it. But suppose I am hallucinating a speckled hen. I know the hallucinated hen has speckles but I cannot know exactly how many. Since access is just a principle of weak self-intimation for obligatoriness, it is vulnerable to the same appeal to indeterminacy.

More concisely, my objection is that if all obligations are knowable, then all obligations are determinate obligations:

All obligations are knowable obligations.

No indeterminate obligation is a knowable obligation.

Therefore, no indeterminate obligation is an obligation.

Anyone who performs this syllogism is in a position to expose any indeterminate obligation as a definite non-obligation.

However, an indeterminate obligation must be one that cannot be known to be an obligation and which cannot be known to be a non-obligation. Therefore, the access principle makes 'indeterminate obligation' an oxymoron. However, the vagueness of language shows that there are infinitely many indeterminate obligations. For example, if Scrooge promises to pay a fair wage to his clerk, what is the minimum he can pay

and still keep his promise? (Further illustrations are provided in Sorensen 1990.)

Mind the ambiguity of 'vague'. In one sense, 'vague' means substandard specificity. An obligation to pay your clerk between \$5 and \$500 per hour is vague in this sense because the interval is so wide. However, it is not vague in the sense that it involves borderline cases. Only vague predicates in the borderline sense generate the sorites paradox. My usage of 'vague' and 'indeterminate' is restricted to this philosophically interesting sense. The attempt to answer the indeterminacy objection by "vaguening up" the level of description relies on an equivocation. Scrooge's "imperfect duty" to pay at least the legal minimum wage is open ended but precise (in its freedom from borderline cases). We can all acknowledge that many vague obligations can be precisified by substituting quotas. But none of us thinks that the precisification is synonymous with the original obligation.

Can the access theorist claim that our real obligation must be a determinate obligation that merely resembles a vague obligation? No, restricting our obligations to determinate obligations just raises the problem to a new level. 'Determinate obligation' itself has borderline cases.

Mixing and iterating determinacy and indeterminacy operators can trigger a feeling of relief by overloading one's parsing ability. The best way of blocking this logical anesthesia is to locate an implication of indeterminacy that does not itself contain the indeterminacy "operator". Refuting that unadorned entailment would refute the obscure entailer:

1. All obligations are knowable. (Assume)
2. It is indeterminate whether I am obliged to vote. (Assume)
3. I cannot know that I am obliged to vote and I cannot know that I am not obliged to vote. (From 2)
4. I cannot know that I am obliged to vote. (3, Simplification)
5. I am not obliged to vote. (1, 4, TF)

If I could know the premises of this proof, I could know the conclusion. But this would conflict with the second conjunct of the third premise. Hence it is impossible for me to know both premises. However, if the access principle is true, it must be knowable. (The knowability of access must be granted by the defender of access because he is presenting himself as knowing that access is true.) That means my indeterminate obligation could not be known to be an indeterminate obligation. The

point generalizes, so the defender of access would be in the peculiar position of saying that there are indeterminate obligations even though agent cannot recognize any of them as indeterminate obligations.

### **C. Dumbing Down Ethics**

A standard motive for public policy secrets is that knowledge of the policy creates perverse incentives. If President Clinton thinks that restoring Aristide to power is not worth the lives of one hundred American soldiers, then publicly announcing so would create an incentive for Haitian attaches to kill Americans.

Knowledge of the publicity principle creates its own perverse incentives. If all of one's obligations were knowable, one could mute the call of duty by diminishing one's cognitive capacity. After all, if I am not obliged to do something, then I am permitted to refrain from doing it. Thus if I can preclude the possibility of knowing that I am obliged to donate some of my inheritance to charity (by burning rather than reading the only copy of a will), then I can ensure that I am permitted to keep the entire inheritance. Simple incuriosity would considerably narrow one's obligations. Exploitation of the

loophole need not be selfish. A concerned parent might keep his children innocent by depriving them of moral knowledge.

The defender of the access principle might first observe that these consequences are not entirely repellent. Genesis 2:16-17 depicts Adam and Eve as innocent until Eve persuades Adam to eat from the Tree of Knowledge. Superiors commonly discourage subordinates from reporting details. Richard Nixon would have completed his presidency if he had a good answer to Senator James Baker's pithy query "What did the President know and when did he know it?".

The natural follow-up is to close the loophole by mandating moral curiosity. For if there is a duty to ascertain our obligations, we can criticize the morally incurious as shirkers. Indeed, after World War II, the Allies hung General Tomoyuki Yamashita on the grounds that he should have known that his troops were committing atrocities in the Philippines. In 1956, the Yamashita ruling was adopted as Army policy. (During the My Lai massacre trials, the American judge rejected this precedent and ruled that Captain Ernest Medina could be held responsible for the crimes of his subordinates only if he had at least indirect knowledge of wrong doing.)

Defending the access principle by means of the curiosity principle requires that the curiosity principle be knowable. For if the curiosity principle were unknowable, it would be invalidated by access principle -- the very proposition it was designed to rescue. But now a higher order loophole opens. If I keep ignorant of whether there is an obligation to ascertain my obligations, I can use the access principle to evade those epistemic obligations even if they exist. To close this meta-loophole, the defender of access must invoke a yet higher order principle to the effect that we have an obligation to learn whether we have an obligation to learn our obligations.

Notice that one's prosecutorial enthusiasm for catching the shirker diminishes as one ascends levels. Suppose one were to show that the shirker violated his tenth order obligation to learn whether he has an obligation to learn whether he has obligation to learn . . . whether he has obligation to read rather than burn the will. Violation of such an esoteric obligation would be a peccadillo compared to burning the will. The feeling of disproportionality is aggravated by the intuition that the will-burner's obligation to donate survives the destruction of its evidence. After all, if the will-burner has a change of heart, he can play it safe by

donating 10% of his inheritance, safer by donating 50%, and can be certain to meet his obligation by donating 100%.

Can the infinite regress be short-circuited by interpreting the curiosity imperative as self-referential? No, because "Everyone has a duty to learn his duties -- including this very duty" is not self-evident. To infer it from itself would be circular. And to infer it from another proposition re-opens the loophole; a shirker can avoid knowledge that the curiosity principle is true by disposing of his opportunities to learn whether it is true.

A genuine solution to this infinite regress might be especially welcomed by ethics instructors. For the curiosity imperative appears to support mandatory ethics courses. However, enrollment in a philosophy course is generally more effective at raising the suspicion that one has overestimated one's knowledge rather than at producing fresh knowledge. Typically, the student is introduced to a variety of moral systems: utilitarianism, egoism, virtue ethics, Kantianism. His inability to eliminate these new alternatives undermines his claim to know principles such as 'One should never deliberately kill innocent people'. Indeed, the student will also confront skeptical challenges from the nihilists and emotivists.

PhD caliber arguments are mounted for the thesis that there are no moral truths and hence no moral knowledge. Thus a student who commences ethical studies out of epistemic duty may conclude that he cannot execute this "duty".

Some of the skeptical pressure can be relieved by reformulating access in terms of probability rather than knowledge:

(Probabilistic access) If one is obliged to do x, then one can be reasonably sure that one is obliged to do x.

This revision will satisfy some fallibilists. However, others will note that sometimes we can only modestly improve our evidence for an hypothesis. Indeed, sometimes the best we can do is to make one alternative slightly more likely than its negation. All other things equal, we should choose the slightly more probable alternative. After all, we should not waste information. But once we admit that the difference between an alternative and its negation can be as small as 1 in a 100 or 1 in 1000, then we must also admit the possibility of there being no evidential difference between the alternatives. Over time, our

evidence for an obligation may flicker in and out -- but the obligation does not flicker in syncopation.

### III. MORAL IMPLOSION

A proponent of access might counter that the object of obligation has been misdescribed: the obligation is to try to pick the better alternative (Prichard 1949, 31-39). If it is always the case that at the moment of action, I can know that my duty is to try to execute my obligation while also knowing the specific option that constitutes trying to execute the obligation, then the access principle is true.

`Try' s-t-r-e-t-c-h-e-s. Like `choose', `try' can be used narrowly to cover just the psychological origin of an act or broadly to encompass the effects this willing. For example, when one imagines someone trying to thread a needle, one pictures the motions involved in trying to insert the thread through the hole. Our imagery focuses on evidence for the mental state rather than the mental state itself. Nevertheless we quickly admit that a totally paralyzed man tries to do things (at least prior to learning of his paralysis).

The ethics of trying focuses on this narrow reading of `try' in order to bring the agent's moral status under his

complete control. It implies that your moral worth would be unaffected if you turned out to be a brain in vat. Widening 'trying' to require some minimal action would be a clumsy compromise. In addition to exposing the agent to moral luck and forsaking those well-meaning brains in a vat, one would still be stuck with all the difficulties wrought by intentions. For the wide readings of 'try' incorporate intentions just as surely as the narrow ones. After all intentions are needed to answer the question 'What were you trying to do by doing that?'

Introspective ethics has been encouraged by Christianity at least since since Augustine's Confessions in the fourth century and has been reinforced by Western individualism. In 1932, H. A. Prichard's "Duty and Ignorance" (reprinted in Prichard 1949) inaugurated an ongoing series of articles on the primacy of subjective duty. But I shall focus on Kant as the most illustrious exponent.

### **A. Objectionable Objects of Intention**

According to Kant, the only thing that is good without qualification is a good will -- a will that aims at duty. Hence morality ends at the edge of intentionality. It does not

continue on into an external world that one cannot control. Kant's retreat has been emulated by subjective utilitarians. They identify right action with choosing the maximum expected utility. Many ethicists welcome this apparent convergence of opposites.

The temptation to conclude that Kantianism and utilitarianism amount to the same thing is encouraged by Kant's insistence that we must try our best (1785, 12-13). Ironically, this maximizing maxim makes Kant vulnerable to the objections against maximizing lodged by satisficers. (New grist for Michael Slote's Beyond Optimizing!) For example, there will be no maximal degree of effort if these degrees are mapped onto fractions less than 1. Such a mapping is natural when one reflects on situations in which it one could have always tried a little harder. For example, a perfectionist is haunted by the realization that each extra hour spent on a painting only improves it asymptotically.

Intuitively, trying half-heartedly is outranked by trying hard and trying hard is outranked by trying one's best. But if one ranks attempts in accordance with how close they come to completion, then it seems arbitrary not to continue the ranking into the realm of successful actions -- and indeed, to rank very

successful actions over moderately successful ones. Yet an ethics of trying must abort this progression toward objective utilitarianism's conclusion that a good will is only of instrumental value. But stopping just before the edge of success creates a problem in specifying the formal object of the moral will. If all obligations are tryings, then trying to meet an obligation means trying to try something. But trying to try what? Trying must aim at something that is not itself a trying. Can one escape the regress by characterizing 'trying to meet one's obligation' as a misleading expression? Perhaps 'obligation' just picks out what one is intending to do. But this would put the situation in the peculiar situation of aiming at a deed that is not itself morally meritorious.

The distinction between reasons for intending and reasons for acting was dramatized in Gregory Kavka's (1983) toxin puzzle. Suppose a billionaire offers you a million dollars if at midnight you intend to drink a vial of toxin tomorrow afternoon. Drinking the toxin would make you ill for a day. Although you would hate to be that sick you would drink the toxin for a million dollars. However, the pay-off will be for intending to drink the toxin. Whether one actually drinks it is irrelevant. At first glance, this seems ideal: your plan should

be to intend to drink the toxin but then not actually drink it. However, knowledge that one will not drink the toxin prevents one from intending to drink it. A reason for intending the action conflicts with a reason to not to carry out the action.

Compare Kavka's toxin drinker to Kant's hero: a man who finds it painful to do his duty but nevertheless performs it out of respect for the moral law (Kant 1785, 17). The hero's ideal outcome is to form the good intention (thereby furnishing all that morality demands) but not do the deed. However, unlike the billionaire, Kant does care about the intentional state of the agent right up to the moment of action. If the hero changes his mind before acting, the backsliding will count morally against the hero. So Kant's hero has a reason to remain resolute to the end.

Nevertheless, I shall argue that the hero's alienation from the deed thwarts his intention. Consider a gapless toxin puzzle. Here, the billionaire insists that you continue to intend to drink. However, he also continues to be indifferent to whether you actually drink the toxin. For example, you will still get the million if the vial of toxin accidentally slips from your lips.

You are still unable to form the million dollar intention. Suppose, for the sake of a contradiction, that you do intentionally drink the toxin. When asked whether you wanted to drink the toxin, you answer "No, all things considered, I intensely preferred not to drink the toxin". When asked whether you drank it out of weakness of will, you reply "No, I drank it because I intended to drink it -- and that intention caused the deed in the normal way". When I ask you whether you are some kind of weirdo (Mele 1992), you respond "No, I am a normal, rational agent -- just the sort of chap that Kant likes".

This pattern of answering is incoherent. Traditional action theory says that you intentionally perform an action only if you perform for it for a reason (Davidson 1980, 264 and Goldman 1970, 76). In addition to violating this principle, you would be violating a more guarded principle: you intentionally perform an action only if any strong reason you have against the act is at least partially offset by a (possibly weaker) reason in its favor.

You can bind yourself to perform a painful action only when you have a reason for the action, not just the intention. If you have insured me against flood damage, then you will hope

that there will be no flood. But if I suffer flood damage, then you have a reason to pay me: your promise to do so. If the insurance policy instead specified that you need only intend to pay me for any flood damage, then the policy does not provide a reason for paying me. (Remember to construe the policy strictly -- don't read in an implied promise to do what you promised to intend to do.)

One may reply that in the gapless toxin puzzle, there is a reason to drink the toxin; that's the only way to get the million! However, if this reply works for the gapless toxin puzzle, it also works for the gappy original. Indeed, this is the solution David Gauthier (1994, 708) advocates for Kavka's original toxin puzzle (and Kavka's earlier paradox of deterrence). Gauthier thinks that an action is prudent as long as it is a component of an overall prudent policy. Since the option of intending but not drinking is infeasible, the rational agent will choose the next most lucrative policy: intending and drinking. Gauthier's position is unpopular. But those who think the million can be won only in the gappy version are forced into the same holistic treatment of rationality. Gauthier may conclude the gapless toxin puzzle provides Kantians with a special reason to side

with him -- along with Gauthier's brand of social contract theory.

Kavka peppered his original presentation of the toxin puzzle with qualifications that eliminated various loopholes. Kavka depicts the hopeful intender as resorting to the final option; you try "to summon up an act of will, gritting your teeth and muttering 'I will drink that toxin' over and over again" (Kavka 1983, 35). But this is also a pseudo-option. Kavka explains why without ever mentioning the time gap feature:

If intentions were inner performances or self-directed commands, you would have no trouble earning your million. You would only need to keep your eye on the clock, and then perform or command to yourself at midnight. Similarly, if intentions were simply decisions, and decisions were volitions fully under the agent's control, there would be no problem. But intentions are better viewed as dispositions to act which are based on reasons to act -- features of the act itself or its (possible) consequences that are valued by the agent. . . . Thus, we can explain your difficulty in earning a fortune: you cannot intend to act as

you have no reason to act, at least when you have substantial reasons not to act. And you (or will have when the time comes) no reason to drink the toxin, and a very good reason not to, for it will make you quite sick for a day.

(Kavka 1983, 35)

Nevertheless, one can see why Kavka succumbed to the temptation of introducing a temporal gap. According to the belief condition for intention, one can intend to do x only if one believes one will do x. This extra premise provides extra fuel for the inference that the toxin drinker cannot make the million.

Although plausible, the belief condition is controversial amongst commentators on intention. Objections have led to diluted versions of the principle such as "One can intend to do x only if one believes it is possible that one will do x". These have also been challenged. The diluted versions also become less helpful to Kavka. After all, there is some chance that you will drink the toxin. You might acquire an independent reason to drink the toxin or merely lapse into stupidity or superstition.

Or you might have erroneous opinions about the nature of intention that will help you drink the toxin. For example,

some people have an impetus model of intention. They think that if one intends to perform an action, then there is momentum toward that action. The longer the intention is in force, the more momentum. In the gapless case the intention is never countermanded. Thus, on the impetus theory, the intention produces action (or at least has as much power to produce action as intentions that are backed by reasons for action). Let us bracket our reservations about the metaphysics of the impetus theory. Let us also suppose that there is an empirical correlation between long standing, uncountermanded intentions and the production of the intended behavior. Toxin drinking would then be akin to the follow-through motion of a bowler. The bowler does not intend to lift his hand above his head after throwing the ball. However, the bowler knows that his intention to throw the ball will cause this hand to continue in an arc that reaches over his head. The lesson here is that only some of the behaviors caused by an intention are intentional actions. The agent's beliefs and desires are used to separate the intentional action from other bodily effects of the intention. This implies that drinking the toxin would not be intentional; it would just be an early part of a stream of behavior triggered by the intention.

Recall Donald Davidson's example of the climber who releases the rope holding his partner because he is unnerved by his belief and desire to release the rope (1980, 79). The climber does not choose to release the rope because the belief and desire do not cause the release in the appropriate manner.

Conceptual error might enable you to form the intention. However, Kavka would eliminate this loophole in the same breath as he forbids the following: implanting intentions by hypnosis, causing yourself to forget relevant facts, hiring hit man to kill you if you fail to drink the toxin, etc. The general spirit of these qualifications suggests that one must form the intention in a rational and informed manner.

The temporal gap in Kavka's original puzzle introduces a red herring. The belief condition distracts us from the role of Kavka's main distinction between reasons for intentions and reasons for actions. And it is this distinction that bedevils Kant and the ethics of trying in general. The logic of intention precludes introverted morality. Obligations must face outward, toward action.

## **B. The Hidden Heart**

A proponent of access is free to limit the ways we can ascertain our obligations. Kant claimed that the limits of human insight testifies to "the wise adaptation of man's cognitive faculties to his practical vocation" (1949, 247). For if we knew more we might act properly merely out of fear or hope because "God and eternity in their awful majesty would stand unceasingly before our eyes". Thus restricting our mode of access increases the opportunity to manifest a good will.

Despite this flexibility, a problem looms for the ethics of trying. Yes, making all obligations internal states ensures that no counterexamples to the access principle will issue from impenetrable regions of the external world. But what about counterexamples that issue from internal darkness? If mental states were weakly self-intimating, then the access principle would face no danger from this quarter. Indeed, it would suffice if just the morally relevant states were accessible at the time of action. However, these morally relevant states were amongst the earliest targets of eighteenth century motivational skeptics. Kant himself warns that

It is in fact absolutely impossible by experience to discern with complete certainty a single case in which the maxim of

an action, however much it may conform to duty, rested solely on moral grounds and on the conception of one's duty. It sometimes happens that in the most searching self-examination we can find nothing except the moral ground of duty which could have been powerful enough to move us to this or that good action and to such great sacrifice. But from this we cannot by any means conclude with certainty that a secret impulse of self-love, falsely appearing as the idea of duty, was not actually the true determining cause of the will. For we like to flatter ourselves with a pretended nobler motive, while in fact even the strictest examination can never lead us entirely behind the secret incentives, for, when moral worth is in question, it is not a matter of actions which one sees but of their inner principles which one does not see.

(1785, 27)

In sum, Kant agrees that we can never know whether we are following the categorical imperative rather than merely acting in accordance with it. Kant may have thought that he could accept motivational skepticism. For one's attention normally focuses on actions and Kant believes the categorical imperative

completely specifies the right action. (No murdering, no lying, no cutting in line.) However, an ethics of trying requires that the right action have the right motivational pedigree. Since the pedigree cannot be known, one cannot know which option constitutes satisfaction of one's obligations.

This point can be illustrated with a modified twin earth thought experiment. In a normal twin earth scenario, we hold constant the individual's entire "narrow" psychology. In this modified scenario, we only hold the accessible portion of the individual's psychology constant. So Will and his doppelganger Twin Will have the same encouraging evidence about their own motivations. However, Twin Will is actually motivated by self-love. Consequently, Kant must appraise Twin Will's conduct as morally worthless.

Tough luck for Twin Will! He had the same evidence as Will. He had the same feelings of earnest effort and the same sense of awe for the starry sky above and the moral law within. But an indiscernible difference between Will and Twin Will makes one of them a satisfier of his obligations and the other an inscrutable pseudo-satisfier. Notice that Twin Will has plenty of moral knowledge. He has even read Kant and has just as much knowledge about the categorical imperative as

Will. However, abstract knowledge is not enough to satisfy the access principle. Will is like someone who knows that he must send power to the blue robot (not the red robot or the green robot) via electricity from the "proper" outlet. The good news is that Will is not responsible if the blue robot malfunctions. The bad news is that Will has no way of telling which outlet is the proper outlet. Fortunately, Will plugged into the proper outlet. Based on the same internal and external evidence, Twin Will plugged into the wrong outlet.

Let's grant that The Foundations of the Metaphysics of Morals specifies behavior in more detail than Spike Lee's movie title "Do the Right Thing". (Kant's movie title would certainly be longer: "Do the Right Thing from the Right Motive".) Let's grant that acting in accordance with the categorical imperative provides evidence that one has a good will. The problem is that this interior morality still does not satisfy the access principle. Indeed, it is reminiscent of Calvinism. A Calvinist believes that God has made good people and bad people. The good are destined to go to Heaven and the bad to Hell. Their actions on earth are causally irrelevant to their fates. However, good works are signs of a good person and so one can raise expected utility by doing

good works. Since people have a psychological tendency to maximize good news, Calvinists have historically been an industrious lot. However, their underlying decision theory (in effect formalized in 1965 by Richard Jeffrey's The Logic of Decision) is undermined by the literature on Newcomb's problem. In any case, the Damned are still inscrutably separated from the Elect by original sin.

Proponents of the ethics of trying may object to my election of Kant as their representative. Isn't Kant an uptight Protestant with a unnatural fondness for noumenal selves? Maybe, but his grounds for doubting how much a person can know about his own motivation are phenomenal. My criticism does not exploit Kantian metaphysics. Anyone who says that the springs of action are always available to introspection is stuck with an unrealistic psychology. Contemporary psychologists who specialize in introspection and the self-attribution of motivation vigorously second Kant's empirical skepticism about our moral self-knowledge (Ross and Nisbett, 1991).

Evolutionary psychology further motivates inaccessible motives. For this economical naturalism builds in a presumption of ignorance. When a species become cave

dwellers, they initially produce offspring with functional eyes. Eventually their eyes atrophy. The reason is not that visual knowledge hurts them. It is just that it fails to help them. Their eyes become superfluous, energy consuming tissues. Thus mutants with simple eye structures are more efficient, and so the species will eventually lose their eyes altogether.

Thus a biologically informed moral epistemologist will apply a presumption of ignorance. He should postulate awareness of morally relevant background facts (and moral knowledge itself) only where it is plausible to think that this kind of knowledge would have made us more reproductively efficient than the alternatives. When weighing the alternatives, we should remember that our basic capacities are tuned to the life of a Pleistocene hunter-gatherer. Over 99% of our history as a species was spent in small, genetically related groups of about 50 people. Only recently have we dwelt in cities filled with strangers.

Mother Nature keeps us informed on a need to know basis (or more precisely, what our hunter-gatherer ancestors needed to know). Our degree of knowledge about our motives will be limited by the usual motley of trade-offs associated

with our knowledge of other aspects of our psychology and our surroundings.

#### IV. EXAMPLES OF UNKNOWABLE OBLIGATIONS

I cannot assert 'I have an unknowable obligation to donate exactly \$284 to Oxfam' because I can only assert what I can appear to know. However, I can assert that 'Saddam Hussein has an obligation to submit to punishment but he is unable to recognize this'. My grounds for saying this are reports about Hussein's messianic intolerance of criticism, his cultivation of yes men, and his pathological self-righteousness. We commonly and coherently ascribe moral blindspots to others by noting how they simply lack the resources to recognize moral facts.

During Clarence Thomas' confirmation hearings, many outraged women complained that "Men just don't get it". 'It' (I think!) refers to how much women are disturbed by sexual forwardness. A man tends to gauge how much a woman will mind a pass from a man by how much he would mind a pass from a woman. Since he would not mind much, he figures she won't mind much. But here the method of empathy misleads. Natural selection has given a different agenda to male and

female desire. Since evolution puts a premium on reproductive success, it has exerted little pressure on men to correct this fallacy and may even abet it (Buss 1994, 146-7).

Given the access principle, it would be inconsistent for women to condemn men for sexual forwardness while excoriating them for being too insensitive to know any better. Women who wish to continue condemning masculine blindspots should reject the access principle and adopt an instrumental view of protest. Continuing to hold the ignorant to their obligations may lead the temporarily uneducable to satisfy their obligations on other grounds. For example, some women realize that many men comply with their wishes in order to be gentlemen or to humor them or out of fear of sanctions. These women hope that the men who act as if they "get it" will eventually come to really get it.

Punishment of the ignorant may be inevitable for all of us at one stage of life. Children cannot learn all of morality by having it explained to them. For example, their natural aggression needs to be promptly curbed or redirected. When a toddler hits his sister, his parents ought to try to make him feel guilty. It is plausible that punishment is needed to trigger concepts of obligation just as exposure to language is needed to

trigger understanding of syntax. To be punishment rather than mere training, the punisher must believe that an obligation has been violated, in this case, the obligation not to hit your sister just for fun. If some obligations can only be known after we are punished for violating them, they must have been in force and yet unknowable at the time we violated them.

"Punish first, explain latter" is a common method of breaking in a new employee. The greenhorn spends his first week "walking on eggshells", struggling for guidance from aloof co-workers. Dissonance theory suggests that this hazing might ultimately produce a loyaler worker. Drill sergeants prefer that recruits learn some rules by humiliating trial and error. Arbitrary rules are unpredictable and so foster reliance on authorities. This heightens obedience and makes the soldier a more malleable fighter.

Knowledge of a particular obligation isn't the only way to reliably satisfy that obligation. Actions correlate with other actions. Prudence normally suffices to ensure that nearly all will avoid self-mutilation. Taste may suffice to prevent cannibalism. Similarly for incest. Most societies have had an incest taboo but lack recognition of the genetic rationale for avoiding mating with first degree relatives. A few societies

have no taboo against incest. They just don't do it.

Evolutionary anthropologists conjecture that these societies do not need the taboo because the children are raised under conditions that decisively activate the negative imprinting that is triggered by being reared together. (Negative imprinting is evident from the absence of intermarriage amongst unrelated Israeli children who are raised together in the kibbutz system.) Individuals in these societies nevertheless have an obligation not to commit incest even if they lack the concept of incest.

#### V. ADMINISTRATIVE ADVANTAGES

These examples of unknowable obligations show that our criteria for ascribing obligations are surprisingly unintellectual. Objective utilitarianism suggests that we should hold people to an obligation when this yields better consequences than not holding them. The administrative advantages of such a policy are familiar from legal discussions of strict liability and the principle that ignorance of the law is no excuse. Obligations are easier to enforce when there is no meta-obligation to verify that the person knew of his obligation. It is difficult to prove a state of mind especially when people are adept at feigning ignorance. People wind up better informed about their

obligations when free of an incentive to remain ignorant. If tax payers knew that the Internal Revenue Service only penalized witting violations of the tax code, then there would be less demand for tax textbooks and consultants. Knowledge of our tax obligations is increased by making knowledge of a tax obligation legally irrelevant. (Notice how 'Ignorance is no excuse' is immune to infinite regress; a tax payer gets nowhere by complaining "But I did not know that ignorance is no excuse".)

The law tempers the principle that ignorance is not an excuse by requiring that prohibitions be disseminated. Illegal parkers can be ticketed only if there is a "No parking" sign. (These signs are posted almost unreadably high in Manhattan because desperate motorists remove or deface signs that are within reach.) One might interpret this dissemination requirement as a concession to the access principle. However, dissemination requirements are more plausibly grounded in efficiency. A small investment in advertising the regulation leads to a large increase in compliance. The dissemination of law tracks the economics of information more closely than the publicity principle.

Mother Nature is as mindful of the economics of information as human institutions. She will design us to be aware of our obligations when awareness yields the best available return. An ethicist can aspire to know more than the bare minimum needed by his hunter-gatherer ancestors. For knowledge that does not earn its keep can often be acquired as a byproduct of abilities that generated useful knowledge. For example, contemporary mathematics issues from a fortuitous combination of an ability to do simple calculations (which many animals have) and language (which only human beings have).

Children who could follow verbal instructions would be more apt to survive to adulthood and have children of their own. Morality may be an elaborate re-programming of this basic cognitive and motivational hardware for parental instruction. No doubt, we are a long way off from discovering the details. But there are leads that warrant speculative elaboration.

## V. PROTO-MORALITY

Egbert Leigh (1977) has argued that the interests of the individual tend to coincide with the interests of his species. For when the divergence of individual interest and group

interest is strong, the species is at a competitive disadvantage from less conflicted species. The same applies to groups within the species. If the group requires strong self-sacrifice, its self-sacrificers will be outreproduced by the more self-centered members. If the individual requires net sacrifice of the group's interests, then the group can only satisfy the demand in a way that weakens it. Hence, there is selective pressure on groups and their constituents to align interests.

The biological function of morality is to facilitate this convergence. If morality did not serve this purpose, it would not exist. This hypothesis only aims at explaining the presence and shape of morality, not to justify it. This concession to the 'is'/'ought' gap is compatible with moral psychology being relevant to justification. Understanding what kind of code a human being can follow narrows the class of what codes they ought to follow.

At some point, human beings achieved a level of intelligence and communication at which it paid to reflect on the relationship between the individual and group. These reflections guided adjustments of the individual with respect to the group and the group with respect to the individual. The

tighter mesh between group and individual interests would enable moral groups to displace amoral ones.

Rivalry would then develop between types of moral codes. The competition could have been through obvious means such as the conquest of pacifists by groups with a more flexible attitude toward violence. However, the holistic nature of fitness ensures that the contribution of a moral code will often be too complicated for armchair anthropology.

Nevertheless, this picture of code competition suggests a correlation between moral thinking and the pace of social change. It also predicts the ascendancy of codes that promote larger and hence more powerful groups. Although our morality must work with a basic social psychology geared to small groups of closely related Pleistocene hunter gatherers, there must have been pressure to retro-fit this mind-set to larger and larger groups of less and less related members. Morality must have become more universal and abstract as the empirical differences between in-group and out-group people were diluted.

Recursive features of morality would be pre-adapted for this transition to big groups of increasingly anonymous individuals. Think of the Confucian hierarchy of obedience:

individual, family, clan, village, state, and nation. Similar nesting occurs for queues, reciprocity, and alliance (a friend of a friend is friend). Structural aspects of morality will be elaborated, content specific aspects will become obsolete (thus the extinction of most taboos).

Recursion is not limited to human beings. Even chickens have pecking orders and path following. Indeed, most elements of morality have a degree of representation in the animal kingdom. Think of how much proto-morality is covered by kin altruism, sympathy, reciprocity, parenting, group protection, territoriality, possession, incest avoidance, rescue behavior, alliance, submission to authority, pegging into a hierarchy, shunning, banishment, punishment, adultery, fighting and greeting rituals. Medieval bestiaries had little trouble using animals to illustrate the virtues of courage (lions), industry (bees), and loyalty (dogs).

Nevertheless, the good deeds of animals strike many people as too disjointed to count as clearly moral. A hen will rescue her chicks if she hears their distress call. But the hen will do nothing if the chick is covered by a bell jar. The hen can see the chick in distress but its rescue program is not activated.

It need not be that the hen is under an illusion that the chick is fine. She may only be unmoved by the chick's plight.

Before we become too smug about the hen's arbitrariness, we should consider social psychology's evidence about human arbitrariness. We grossly over-estimate the uniformity of individual behavior by neglecting potent situational variables. If a starving baby is placed on a homeowner's doorstep, then he engages in rescue behavior. If the homeowner is merely informed of a starving baby in a far away place, then the homeowner will do nothing.

At least portions of moral behavior can be explained as a patchwork of programs. Some programs may be inflexible like the rigid sequence of acts constituting a yawn. Others may easily accept grafting and pruning. However, we need only follow the instructions of a moral code in the way that a computer executes its program (Cummins 1987). Programs can (but need not) be followed in a conscious way. Alan Turing himself was the first computer chess player: for want of a machine, he laboriously followed the instructions himself. (And lost to a duffer!)

Consciousness is wasted on routine matters. This explains why we willingly automate altruism. Instead of

writing a check each month for famine relief, we have charitable contributions deducted by a payroll program. Set it and forget it! The older but still dominant way of mechanizing morality is by habit. What begins in reflection and effort, gradually becomes unthinking routine. Unconscious processes introduce shortcuts into this mushrooming mass of good nature. Our robotic superego acquires layer after layer of new habits that compress and deform the old layers. Digging through the sediment requires special skills. Not all of it can be revealed by mental geology.

Those who think this increasing automation decreases the moral value of the deeds are fetishizing consciousness. True, we do admire the reflective defiance of Socrates. But we also admire the modesty of Cincinnatus and the impulsive courage of Russians quelling the Chernobyl nuclear disaster (Driver 1989). People take pride in what they think about and in what they don't need think about.

When mechanisms for meeting our obligations are highly reliable, there will be little need for us to be aware of those obligations. As the pressure to be aware dissipates, so does the capacity to become aware. Consciousness, after all, is largely devoted to trouble-shooting. Our attention is drawn to

the novel and problematic, not to the routine and satisfactory. The pattern for moral knowledge should thus follow the pattern of other knowledge.

## VII. WHY IGNORANCE SEEMS RELEVANT

Ignorance is more important in normative matters than factual matters because the observer effects are larger. But subjective utilitarians have mislocated these effects in the agent. The real subjectivity flows from those who wish to influence the agent's behavior. Since the agent sometimes wants to influence his own behavior, his beliefs can also be relevant on the same grounds. However, his beliefs have no special status.

This point can be made by contrasting rightness with obligation. I think students initially accept the utilitarian definition of 'right action' because it is indeed an analytic truth. However, they mistakenly abandon the definition because of "counterexamples" that assume we are obliged to do what is right. This initially compelling assumption is less plausible if we reflect on the intimate connection between obligation and blame.

To say that someone is obliged to do something is to be prepared to blame him if he fails to meet his obligation.

Sometimes we are also prepared to praise compliance. For supererogatory acts -- ones "above and beyond the call of duty" -- we are only prepared to praise but not blame. Finally, there is a neutral realm in which we refrain from blame and refrain from praise.

We can now ask when it is right to praise and blame? (Notice that this question differs from 'When are we obliged to praise and blame?'.) The utilitarian answers that we should praise and blame in accordance with the consequences of the praise and blame. However, this answer must be qualified because there is a belief component to praise and blame. 'You did not steal my bicycle but I blame you for stealing my bicycle' has the same absurdity as G. E. Moore's 'It is raining but I do not believe it'. Both sentences are unbelievable. If Kavka's billionaire offers to donate a million dollars to Oxfam if I blame you for stealing the bicycle, then I cannot make the million. Of course, I can make the million if I only need to say that you stole the bicycle. But that's not the deal.

Retributivists get a false lift from the belief component of blame. They criticize utilitarianism for sometimes instructing us to blame the innocent. They present this as a moral absurdity but the root absurdity plays off of the nature

of belief. Since right actions must always be possible actions, the utilitarian agrees that it is never right to blame the innocent. Since punishment implies blame, utilitarianism also implies that it is never right to punish the innocent.

Blame creates the kind of phenomenology that deontologists love because it cannot be directly affected by information about its consequences. However, this is no more damaging to utilitarianism than a parallel observation for other propositional attitudes. Of course, utilitarianism does imply that it is sometimes permissible to insincerely accuse others of wrong-doing. But this is no more damaging than the utilitarianism's implications about the occasional rightness of lying, cheating, and stealing.

Just as one can believe that one has some false beliefs, one can believe that some of one's blamings are mistaken. Thus the utilitarian can punish the innocent in the statistical sense. The law sets standards of evidence high to reduce the number of people who will be falsely convicted. "Better that ten guilty men go free than that one innocent man be punished." But what if the ratio is a hundred to one or a thousand to one or a million to one? Unless one is willing to abandon punishment altogether, one must accept some

punishment of the innocent. Once the retributivist has conceded the practicality of drawing a line somewhere, he has no qualitative superiority on the issue.

Although we lack any direct control over blame, we can indirectly control it by adopting standards of evidence, fostering habits such as role reversal, and so forth. Utilitarian thinking influences these indirect measures. It says that we should foster blame that produces the best consequences. Since blaming and being blamed is painful business, blame must be justified as a necessary evil. Blame functions mainly (but not solely) as a penalty. If one threatens to impose the penalty, one may gain compliance without inflicting the evil. Hence, we usually have an obligation to let people know what we are prepared to blame them for. That is, we usually have an obligation to let people know what we think their obligations are. Since the distinction between what I think your obligation is and your obligation is pragmatically equivalent for me, this has the same practical force as an obligation to inform you of your obligations. Thus objective act utilitarianism implies a rule of thumb that resembles the access principle.

However, it is only a rule of thumb. Deterrence is the chief redeeming good of punishment but not the only. Punishment has pedagogical value. It also works non-cognitively by triggering moral development and moral conversions. There are also values noted by the emotivists. Blame advertises one's moral views, strengthens one's commitment, and consolidates fellow blamers into a coalition. Greater sensitivity to the social psychology of blame would bring a utilitarian theory of punishment to a more complete state. However, the self-reflexive nature of instrumental blame suggests that a utilitarian would always have an incomplete account of just when he should blame. A complete account would require the utilitarian to blame only after considering the consequences of blaming -- plus the consequences of that higher order prediction -- plus the consequences of that still higher order prediction. So even if the utilitarian had an unlimited research budget, he would need to resign himself to unknowable obligations. But so should we all.

\*Ted Sider has graciously permitted me to become a moving target. I have revised in response to his initial reply and he has revised in response to this revision. I am also grateful to Walter Sinnott-Armstrong and Sigrun Svavarsdottir for detailed criticisms and tips.

#### REFERENCES

- Baier, Kurt (1958) The Moral Point of View, Ithaca: Cornell University Press.
- Buss, David (1994) The Evolution of Desire New York: Basic Books.
- Chisholm, Roderick (1942) "The Problem of the Speckled Hen", Mind, 51/204: 349-52.
- Cummins, Robert (1987) "Programs in the Explanation of Behavior", in Scientific Knowledge ed. Janet Kourany, Belmont, California: Wadsworth.
- Davidson, Donald (1980) Essays on Actions and Events, Oxford: Clarendon Press.
- Driver, Julia (1989) "The Virtues of Ignorance", Journal of Philosophy 86: 373-384.
- Gauthier, David (1994) "Assure and Threaten" Ethics 104 (July 1994): 690-721.

Goldman, Alvin (1970) A Theory of Human Action, Englewood Cliffs, New Jersey: Prentice Hall.

Kant, Immanuel (1785, 1969) Foundations of the Metaphysics of Morals, trans. Lewis White Beck, ed. Robert Paul Wolff, New York: MacMillan.

\_\_\_\_\_ (1949) Critique of Practical Reason and Other Writings in Moral Philosophy, trans. Lewis White Beck, Chicago University Press.

Kavka, Gregory (1983) "The Toxin Puzzle", Analysis 43: 33-6.

Leigh, Egbert (1977) "How does selection reconcile individual advantage with the good of the group?", Proceedings of the National Academy of Sciences 74: 4542-6.

Mele, Alfred R. (1992) "Intentions, Reasons, and Beliefs", Philosophical Studies 68: 171-193.

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

Prichard, H. A. (1949) Moral Obligation, Oxford: Oxford University Press.

Rawls, John (1971) A Theory of Justice, Cambridge: Harvard University Press.

Ross, Lee and Nisbett, Richard (1991) The Person and the Situation, Philadelphia: Temple University Press.

Schellenberg, J. L. (1993) Divine Hiddenness and Human Reason, Ithaca, New York: Cornell University Press.

Sidgwick, Henry (1907) The Methods of Ethics, 7th edition (London: Macmillan, 1907).

Slote, Michael "Morality and Ignorance", Journal of Philosophy 74 (December 1977) 745-67.

Sorensen, Roy (1990) "Vagueness implies Cognitivism", American Philosophical Quarterly 27: 1-14.

\_\_\_\_\_ (1991) "Moral Dilemmas, Thought Experiments, and Conflict Vagueness", Philosophical Studies 63: 291-308.

Williams, Bernard and Smart, J. J. C. Smart (1973) Utilitarianism: For and Against, Cambridge: Cambridge University Press.

Williams, Bernard (1976) "Moral Luck" Proceedings of the Aristotelian Society supplementary volume 1: 115-135.