Forthcoming in Philosophy and Phenomenological Research

CHARITY IMPLIES META-CHARITY

"As I am, so I see." -- Ralph Waldo Emerson

"It is irrational to believe others are irrational". I ungratefully said that to a confidant who asserted that I was negotiating with a fool. I now wonder whether I was the real fool. If I believe my friend is irrational (in light of his attribution of irrationality to the recipient of my offers), then my epigram implies *I* am irrational. To avoid the implication that I am irrational, I must not believe anyone to be irrational. But then my epigram also forbids me from believing that *someone else* believes someone is irrational. I must instead believe that the non-existence of irrationality is common knowledge!

1. Volcanic rationality

Can I simply repudiate my epigram? I hesitate because the epigram is a consequence of the principle of charity. Roughly, this interpretive principle states that all agents are rational agents. The standard reasoning behind the principle is *a priori*: there is a conceptual connection between regarding someone as an agent and viewing his beliefs and desires as forming a coherent system that makes his actions intelligible. Since errors about central *a priori* truths indict one's rationality, failure to believe the principle of charity would be irrational. Consequently, charity implies all agents believe all agents are rational. Charity iterates. Since meta-charity would also be a central *a priori* truth, meta-charity implies meta-meta-charity. And meta-meta-charity implies meta-meta-meta-charity. And so on. Therefore, if the principle of charity is true, then it is common knowledge. So is meta-charity. And meta-meta-charity. And so on.

Developers of doxastic logic appear to accept the entailment thesis as a matter of course. Their models of belief normally preclude inconsistent belief and so preclude the attribution of inconsistency to *other* believers. (For a detailed specimen, see Leonard Linsky's 1968 analysis of Jaakko Hintikka's logic of belief.) Most doxastic logicians try to smooth over the conflict

between their logics of belief and ordinary belief by either restricting their claims to *ideal* thinkers or by postulating a separate sense of `belief' for which their doxastic logic holds.

My thesis that charity implies meta-charity does not rely on idealization or the postulation of a special sense of `belief'. The standard arguments for charity are directed to the everyday concept of belief. My thesis is that these arguments about ordinary belief support meta-charity. Proponents of the principle of charity concede that their premises are potent stuff. They tacitly assume that these premises can be titrated to yield moderate conclusions.

I contend that charity is too volatile to achieve this measured effect. If heavily diluted, charity cannot perform functions such as guiding translation and explaining actions. If agents are given a high enough dose to do the job, charity snowballs into common knowledge of rationality. Like chemotherapeutic drugs, charity has no comfortable intermediate dosage.

There is a rabble of arguments for the principle of charity. Part of the discordant variety is due to lack of coordination as to what counts as the principle of charity. But even those who converge on the formulation of charity diverge on what premises yield a sound argument for that conclusion. For instance, many proofs of charity focus on its *a priori* status as a first principle of interpretation. Yet a leading proponent of charity, W. V. Quine, has famous reservations about the distinction between *a priori* and *a posteriori* truths.

Yet other champions of charity are surprisingly indifferent about whether there really are beliefs. The instrumentalist Daniel Dennett maintains we attribute beliefs and desires with the same ontological light heartedness that physicists assign centers of gravity. Dennett's "intentional stance" is a predictive framework in which rationality plays a defining role. Anyone who superimposes the longitudes of desire and the latitudes of belief is already attributing rationality.

Realists about beliefs and desires distinguish between literal attributions of rationality and fictive attributions (say to a thermostat). But Dennett's only criterion is predictive success. Since he is bullish on the predictive power of the intentional stance, he treats animals, computers, and even thermostats as rational agents.

An instrumentalist is free to doubt whether there really are any beliefs. If there are no beliefs, then there are no agents and the entailment thesis `Charity implies meta-charity' is vacuously true. Thus I count all eliminativists about beliefs (such as Chuchland 1979) as degenerate co-subscribers to `Charity implies meta-charity'. But Dennett is not merely an ally on a technicality. His account of charity furnishes substantive support for the entailment thesis. After all, Dennett's theme is that the attribution of rationality is constitutive of the framework that employs beliefs and desires. His defence of charity is conceptual, not metaphysical.

Bear in mind that I am not trying to justify the principle of charity. I am merely defending an entailment thesis. I accept the entailment thesis in both its subjective and objective forms:

Objective form: `All agents are rational' implies `All agents believe all agents are rational'.

Subjective form: Anyone who believes that all agents are rational ought to believe that all agents believe that all agents are rational.

Some might suggest scenarios in which one form of the entailment thesis is true and other is false. Suppose the objective version is true but too subtlely true for us to recognize -- even by those who believed that all agents are rational. Since `ought' implies `can', the subjective form of charity would then be false. Some might suggest the opposite kind of scenario in which the subjective form of the entailment thesis is true but the objective form is false. Suppose people sleepwalk through their applications of charity: If everybody interpreted one another as being rational without anybody realizing they were engaged in this practice, the objective form of charity would be false. The subjective form of charity would be a vacuously true conditional.

Could we really overlook our own practice of treating agents as rational agents? Admittedly, people can be unaware of what is happening right under their noses. The notoriously absent-minded David Hilbert once asked the physicist James Franck: "James, is your

wife as mean as mine?". Puzzled, James Franck asked the mathematician "Well, what has your wife done?". Hilbert replied "It was only this morning that I discovered quite by accident that my wife does not give me an egg for breakfast. Heaven knows how long that has been going on." (Mehra 1973)

Hilbert cannot be as oblivious of his use of charity as he is of the contents of his breakfast. Hilbert interprets Franck's reply as a request for further information that would help Franck understand Hilbert's question. This meta-interpretation requires Hilbert to picture Franck as attempting to attribute beliefs and desires that would make sense of Hilbert's first question. Speakers meet their hearers half-way. In choosing supplemental remarks, speakers presuppose that the hearers are using a rationality requirement to test viable hypotheses as to the speaker's meaning. Thus Hilbert's own use of charity is reflected back in his interpretation of Franck's interpreting. The symmetry of meta-interpretation is a mirror displaying one's own interpretive practices.

Yet people do sincerely deny that they subscribe to even a weak form of charity. Proponents of charity are paternalistically enrolling others in their organization. I am aware of the trade-off between liberties that lead unions to demand a "closed shop" in which all workers are compelled to be members of the union. My point is that if we accept this kind of heavy-handed attribution of charity, we should also accept (from that same heavy hand) an attribution of meta-charity.

Some of my arguments defend the objective reading of `Charity implies meta-charity'. Others defend the subjective reading. The basic theme behind both kinds of arguments is that the reasoning for charity also yields meta-charity. Since `Charity implies meta-charity' spirals up to common knowledge of the principle, the entailment thesis attributes much more content to charity than has been previously assumed. Opponents of charity are free to embrace the entailment thesis and use it in a new *modus tollens* refutation of the principle of charity.

2. Natural and theoretical appeals to meta-charity

When I was a boy, I methodically searched for forgotten change in mechanical coin returns. Occasional triumphs in phone booths led me to view my coin checking hobby as a possible escape from school. My father had explained that the purpose of attending school was to get a job and the purpose of working at a job was to get money. Since the mechanical coin returns had money for the taking, I argued that school was an unnecessary intermediate step. My father replied that I could never make as much money from coin returns as from a regular job. I asked him how he knew this. Had *he* tried making a living from searching coin returns? My father answered that he did not need to try. If grown men could make a living from coin returns, then they would quit their jobs and compete for the coins. That would reduce the number of coins to a level that no longer made coin checking a viable occupation. That's why you see only children and vagrants going from coin return to coin return. So without checking, he knew there was not enough money in those coin returns. And so did all the other men because the same reasoning is available to them. That is why children must attend school.

My father's *a priori* explanation tacitly employs charity and meta-charity. He assumes that other men are rational agents maximizing their expected return on their labor. If there were plenty of cash for the taking, these men would seek the easy money rather than work at their regular jobs. But they do not even bother to look for free money. They realize that other men would have already taken it. So the working men, in addition to being rational, view each other as being rational.

Notice that my father's explanation also contains the seeds of an economic theory of forgetfulness. People cannot directly forget but they can refrain from taking measures needed to remember. When coins were worth more, people must have been more assiduous in checking for their change. In the 1970s, operators at Milwaukee's processing plant estimated that each ton of garbage contains $2.38 in change (Alexander 1993, 23). I predict that the amount has climbed with inflation. People who foresee that they will throw money away do not intend to do so. They just do not think it worthwhile to police their habits. Many apparent irrationalities can be construed as lapses of agency -- over which we have some control. Many "foolish deeds" are

just mechanical pieces of behavior. Since attention is a scarce resource, we run on auto-pilot. Insofar as we proceed as unthinking machines, we are not open to appraisal as agents.

My attitude toward school changed after a university course in philosophy. But my neglect of commercially oriented courses eventually left me more and more concerned about how I would earn a living. One possibility was investment in the stock market. Busy doctors and lawyers must have little time to research their investments. Since I anticipated a future of idleness, I thought I could out-study them. But then I discovered a theory that persuaded me to abandon this plan to buy low and sell high. According to the efficient market hypothesis (EMH) it is impossible to make substantially profitable predictions of stock market prices with publicly available information. For if the substantial profits could be made, other competent predictors would use the same information and bid up the price of the stock. The stock market is systematically over-researched. The practical consequence is that one cannot do better than a random investment strategy in which one buys and sells stocks without any research. Recognition of the futility of research leads one to follow a buy and hold policy to minimize transaction costs. Since those who deny the EMH will have a motive to switch stocks, they will incur transaction costs that will make their profits lower than the investor applying the EMH. Academic economists have conducted many statistical studies that confirm this prediction. However, most proponents of the EMH are quick to emphasize that the efficiency of the market would disappear if everyone believed the EMH. For then no one would be doing the research. Thus the EMH appears to boldly imply that rational, well-informed people disagree with it. The apparent implication is that unbelievers are rational but regard most investors as irrational. Consequently, the EMH appears to imply charity while precluding meta-charity.

If the EMH really implies that it will be rejected by a great body of rational, well-informed people, then we should lower our confidence in the EMH (Sorensen 1983). Experts count. The mere fact that the EMH predicted disagreement with itself does little to cushion the blow. Current forms of creationism predict that professional biologists will disagree with creationism but that does not neutralize the evidential impact of the disagreement. The

theoretician cannot cancel the evidential impact of expert opinion by pre-emptively adding the conjunct `and the experts will disagree with my hypothesis'. A theory that predicts people will present strong evidence against it is a theory that illustrates how a hypothesis can be *dis*confirmed by its own correct predictions.

I now wonder whether the proponent of the EMH should concede that the EMH implies that there will be massive, well-informed rejection of the EMH. The proponent of the EMH could instead portray those who attempt to make profitable predictions in the way my father portrayed children and vagrants. Strictly speaking, the EMH concedes that one can make predictions that systematically outperform random investment. The EMH merely denies that the predictions can outperform the market enough to make the effort worthwhile. What is worthwhile varies from individual to individual. Men who patrol beaches with electronic metal detectors intrinsically enjoy the activity. It is a hobby. Similarly, treasure hunters at sea enjoy the historical detective work and get a gambler's thrill from the possibility of a huge prize. Those who research the stock market could be pictured the same way. Even someone who believed the EMH might "play the stock market" because he was risk loving or enjoyed the life of the short-term investor. Thus the EMH does not really imply disagreement with itself. The disagreement is inferred from the reports of those who try to make profitable predictions. These individuals would deny the analogy with children and vagrants. They say their predictions will earn them an expected return that is superior to regular jobs.

A firm proponent of the EMH and meta-charity might deny that these ambitious investors really believe what they say. Compare them to people who buy lottery tickets. These people often claim to be lucky or savvy. The economist is reluctant to take these assertions at face value. Instead of interpreting the lottery participants as disagreeing with the standard analysis that a lottery ticket is a poor investment, the economist postulates desires that make purchasing lottery tickets rational for thrill seeking purchasers. Economists interpret people as rational and scold those who interpret lottery participants as irrational.

3. The rubber and glue effect

Recall the child's verbal defense: "I am rubber and you are glue, Whatever you say bounces off of me and sticks to you." This incantation is superfluous when one is sincerely but falsely accused of an error. On June 15, 1992 there was a spelling bee at the Rivera Elementary School in Trenton, New Jersey. Vice-President Dan Qualye officiated. When the twelve year-old William Figueroa wrote *potato* on the blackboard, Qualye hinted that the word was missing "the little bit at the end". After more gentle prodding, the boy reluctantly added an *e* at the end to please the Vice-President. Quayle's mis-correction was embarrassing. If you falsely believe someone else mis-spelled that word, then you yourself have mis-spelled it. Specific mis-accusations of mis-spelling boomerang.

The same poetic justice is exacted in false accusations of probability mis-calculations. In September 1991, Marilyn vos Savant published a solution to the Monty Hall problem in Parade magazine. Many mathematicians (among thousands of others) falsely accused Marilyn vos Savant of committing a probability fallacy. The mathematicians suffered professional embarrassment because Marilyn being right meant they were wrong. A mistake at the meta-level implies a mistake at the object level.

Given the *a priori* status of the principle of charity, Karl Popper also commits a logical error when he says that "the method of applying a situational logic . . . is not based on any psychological assumption concerning the rationality (or otherwise) of `human nature' (1957, 97). To regard an analytic truth as only contingently true is an *analytical* error. Hence, those who believe that the principle of charity is made true by the meaning of `belief' must ascribe an analytical error to Carl Hempel on the grounds that Hempel (1962, 12) regards the principle of charity as an empirical issue.

Given that people have the rationality imputed by the principle of charity, all uncharitable attributions would be irrational. Misinterpretations of the rational as irrational are themselves irrational. Therefore, loyalty to the principle of charity should generate vicarious fear

of the rubber and glue effect. To avoid attributing irrationality, the charitable interpreter must avoid attributing uncharitable interpretations to others.

This higher level charity is hard to reconcile with what ordinary people say. Why does the frustrated motorist accuse others of irrationality if he does not believe the charge and does not believe others will believe it?

One answer characterizes the angry attribution of irrationality as street theater. In the aftermath of a traffic accident, motorists also accuse each other of being deaf and blind. If the accuser really believes that the accused is deaf, then why does he deliver the charge orally instead of in writing? If the accuser believes the accused is blind, then why does he point to the puddle of green anti-freeze to support his contention that his radiator has been ruptured? And if the accuser thinks the accused is irrational, then why does he expect him to infer that ruptured radiators are expensive to repair? In truth, the accuser does not believe his accusations and does not believe that others will believe them. That is why the motorist's false accusations are not lies.

We pressure each other principally by the *act* of confrontation, not the cogency of what is said in heated conversation. Social psychologists will predict who capitulates and who compromises by assessing the circumstances in which the dispute takes place, not by the plausibility of the accusations. The motorist’s curses may have illogical content without indicting his rationality. We must interpret the rantings of the irate motorist as we interpret poets.

Just as strong emotion can be expressed by violating norms (smashing vases), strong emotion can also be conveyed by ostentatiously violating conventions for cooperative conversation. We justifiably infer anger when people pollute their channel of communication with conspicuous falsehoods, irrelevancies, loudness, and violations of turn-taking etiquette. Characterizing this attack on the joint good of dialogue as "irrational" is often itself a rhetorical device. `Irrational' is a pejorative term that can be deployed to keep the peace.

4. Maximizing agreement fosters meta-charity

N. L. Wilson (1958-9) introduced the principle of charity as a requirement to maximize assignments of `true' to translated utterances. If Wilson meant this to apply strictly to just what was said, the principle would come to grief with our frequent use of metaphor, sarcasm, deception, and linguistic play. When my six year old son is dragooned into a shopping trip, he eases his boredom by playing the "opposite game". Players must say the opposite of what they intend. Although we never explain the game to onlookers, they quickly catch on. These charitable outsiders penetrate our language game by maximizing assignments of *false* to our initially enigmatic sentences.

Wilson is more charitably interpreted as maintaining that we maximize the attribution of true *beliefs*. As Donald Davidson (1984b, 169) observes, this means that the interpreter maximizes *perceived* truth. So in effect, the interpreter winds up maximizing *agreement* with those he is interpreting. This creates direct pressure to construe other agents as fellow subscribers to the principle of charity.

There is also indirect pressure. Assume there are two interpreters and a single interpretee. Two variants of this three person situation are unproblematic. If there is one interpreter, then he will interpret both parties as, by and large, agreeing with each other, and as also agreeing with the interpreter. If there are two interpreters and each follow the principle of charity, then there is again no problem because each interpreter will strive for a three way agreement. But now for the tricky case. Suppose Mr. Charity and Mr. Uncharity are interpreting the same man. Only Mr. Charity aims at maximizing his agreement with the third man. Barring an accidental convergence, Mr. Uncharity will have substantial disagreements with the third man. Consequently, Mr. Charity will be forced to attribute many false beliefs to Mr. Uncharity, in particular, false beliefs about the third man's beliefs. The scale of the false belief attribution will be large. At least it should be perceived as large by those who subscribe to the principle of charity. They think the principle of charity plays a critical role in belief attribution.

The error caused by uncharity would frustrate Mr. Charity's effort to maximize agreement with Mr. Uncharity. Since Mr. Charity wishes to be *universally* charitable, his only

recourse is to deny that there really are uncharitable people. Mr. Uncharity *cannot* exist! In other words, Mr. Charity will assimilate the situation to the previous unproblematic situation in which there are two interpreters each of whom use the principle of charity.

5. Davidson on the indispensability of charity

If `Charity implies meta-charity' is false, then a person could be rational and believe that some other agents are not rational. That is, he could rationally attribute beliefs to others without the help of the principle of charity. This conflicts with the common characterization of charity as a mandatory preparatory condition for any belief attribution whatsoever. Davidson emphasizes that the attribution of belief and the attribution of meaning are inseparable projects. Both must be done together when attempting to understand speech behavior:

> Since knowledge of beliefs comes only with the ability to interpret words, the only possibility at the start is to assume general agreement on beliefs. We get a first approximation to a finished theory by assigning to sentences of a speaker conditions of truth that actually obtain (in our opinion) just when the speaker holds these sentences true. (Davidson 1984a, 196)

According to Davidson, someone who does not presume massive agreement cannot understand speech behavior. So if Davidson interprets someone as understanding another person's speech, Davidson will attribute the principle of charity to him.

> Since charity is not an option, but a condition of having a workable theory [of radical interpretation], it is meaningless to suggest that we might fall into massive error by endorsing it. Until we have successfully established a systematic correlation of sentences held true with sentences held true, there are no mistakes to make. Charity is forced on us;

> – whether we like it or not, if we want to understand others, we must count them right in most matters. (Davidson 1984a: 197)

In addition to being a necessary starting point, agreement is needed to shape disagreement. For shared beliefs are needed to ensure that both parties are talking about the same thing.

> . . . a belief is identified by its location in a pattern of beliefs; it is this pattern that determines the subject-matter of the belief, what the belief is about. Before some object in, or aspect of, the world can become part of the subject-matter of a belief (true or false) there must be endless true beliefs about the subject-matter. False beliefs tend to undermine the identification of the subject-matter; to undermine, therefore, the validity of the description of the belief as being about that subject. And so, in turn, false beliefs undermine the claim that a connected belief is false. To take an example, how clear are we that the ancients – some ancients – believed that the earth was flat? *This earth*? Well, the earth of ours is part of the solar system, a system partly identified by the fact that is a gaggle of large, cool, solid bodies circling around a very large, hot star. If someone believes *none* of this about the earth is it certain that it is the earth that he is thinking about? (Davidson 1984b, 168)

This passage illustrates how the principle of charity is sensitive to the distribution of false beliefs, not just their quantity. If charity were simply a matter of maximizing true beliefs, then one might hope to make room for a patch of irrationality by concentrating the small amount of disagreement into one area. But if each false belief must be understood against a background of local agreement (to establish the subject-matter of the false belief), then there cannot be any ghettoes of false beliefs.

Davidson cannot take himself to disagree with a genuine abstainer from charity. There could only be a verbal dispute because the non-practitioner of charity would not be able to

muster the agreement needed to disagree with anyone. In “Thought and Talk” Davidson maintains that “a speaker must himself be an interpreter of others”. Since all interpreters must use the principle of charity, Davidson himself is a subscriber to meta-charity: all agents believe all agents are rational. The thesis of “Thought and Talk” is that a “creature cannot have thoughts unless it is an interpreter of the speech of another”. Consequently, Davidson appears committed to the even stronger position that all creatures capable of thought believe all agents are rational. Davidson does not indicate whether he believes all agents believe meta-charity.

One can subscribe to meta-charity without accepting the entailment thesis that charity implies meta-charity. Consider David Lewis once characterized charity as a platitude.

> I have said, rather loosely, that the fundamental principles of our common-sense theory of persons implicitly define such concepts as belief, desire, and meaning. Actually, I would like to claim something stronger: that the implicit definitions can be made explicit, and that the explicit definitions so obtained would be analytic. If so, then our constraining principles would themselves have a status akin to analyticity: Karl might have no beliefs, desires, or meanings at all, but it is analytic that if he does have them then they more or less conform to the constraining principles by which the concepts of belief, desire, and meaning are defined. (Lewis 1983, 112)

Since platitudes are universally believed, Lewis must believe that others believe the principle of charity. However, a platitude need not be recognized as having the status of mere platitude. Recall how John Locke scorned those who regarded `Each thing is identical to itself’ as a learned insight. If the principle of charity were merely a platitude, then it could be universally believed without any of its subscribers being aware of anyone else’s agreement with the principle. Lewis’s insight that charity is a platitude would generate *his* belief in meta-charity. But Lewis would not thereby be in a position to attribute meta-charity to his co-believers in charity. Still less would he be able to infer that the principle of charity is common knowledge.

Nevertheless, Lewis's allegiance to charity should make him receptive to the entailment thesis. John Locke's scorn is uncharitable. If a statement is an obvious analytic truth, then it is irrational to embrace it as arcane wisdom.

Whoops! There I go again. If I accuse John Locke of irrationally attributing irrationality to others, then I jeopardize my subscription to charity. The principle of charity instructs me to retract my charge that John Locke uncharitably interpreted anyone.

Although meta-charity does not imply the entailment thesis, it is enough to make an enigma of any disagreement about whether the principle of charity is correct. The meta-charity advocate must attribute belief in charity to his adversaries. So how can he account for all the scholarly criticism of charity? It is possible to disagree with someone who agrees with you -- if he inconsistently both believes and disbelieves your thesis. For instance, mildly superstitious people who fear the number 13 seem to both believe the number is unlucky and to share the common sense belief that all numbers are harmless. The point of arguing is to extinguish his disbelief in your thesis that 13 is unlucky, not to ignite belief that 13 is harmless (because you think the superstitious man already believes 13 is harmless). G. E. Moore seems to treat his adversaries this way. Moore assumes the skeptic really believes he knows Moore is holding a pen, the solipsist really believes that he is arguing with another person, and so on. When it comes to non-technical issues, Moore assumes he is preaching to the choir -- albeit an inconsistent choir. The proponent of meta-charity cannot use this pattern of explanation to account for the controversy over the principle of charity. ("Come now, you know very well that all agents are rational, and you know very well that all agents believe all agents are rational, and that indeed, the whole matter is common knowledge.") For that hypothesis presents his adversaries as inconsistent; they both believe and disbelieve the principle of charity. Therefore, as an advocate of the principle of charity, the believer in meta-charity must deny that there is any genuine controversy about it. Indeed, he must hold that there is a universal consensus about the principle of charity. He must dismiss disagreement about charity as analytically impossible!

6. Can one choose whether to be rational?

The classic question `Why should I be rational?' presupposes that I have a choice whether to be rational. Obviously I have a choice whether to remain an agent. When I decide to nap, I rationally choose to enter a state in which I cannot choose. But can I choose to enter a state in which I am an agent but not a rational agent?

Given charity, the only kind of agency is rational agency. Thus there would be no *objective* freedom to choose between rational agency and some other form of agency. Given meta-charity, there would be no *subjective* freedom because one would recognize that all agents are rational agents. Out goes the existential freedom extolled by the underground man:

> there is one case, one only, when man may consciously, purposely, desire what is injurious to himself, what is stupid, very stupid -- simply in order to have the right to desire for himself even what is very stupid and not to be bound by an obligation to desire only what is sensible. Of course, this very stupid thing, this caprice of ours, may be in reality, gentlemen, more advantageous for us than anything else on earth, especially in certain cases. And in particular it may be more advantageous than any advantage even when it does us obvious harm, and contradicts the soundest conclusions of our reason concerning our advantage -- for in any circumstances it preserves for us what is most precious and most important -- that is, our personality, our individuality. (Dostoevsky1918, 72)

Given charity, the underground man is condemned to rationality. Given meta-charity, he cannot even hope for irrationality. For the underground man, the skeletal architecture of belief-desire explanation is a cage of rationality.

7. Can irrationality be faked?

Introductory textbooks on game theory warn the reader not to infer irrationality from "inconsistent" behavior. The baseball pitcher Brian Anderson is better at throwing fastballs than sliders. But if he always threw his best pitch, batters would know what to expect. Therefore, Anderson randomizes, throwing a mixture of pitches to keep the batter uncertain.

Game theorists concede that irrationality is sometimes systematically more rewarding than rationality. (Shakespeare too: "Though this be madness, yet there be method in't" Hamlet II, ii.) In the game of chicken, California teenagers drive cars toward each other until one of them swerves to avoid the head-on collision. If a driver is persuaded that his adversary is irrational, then he should swerve. Realizing this, drivers feign irrationality. But if charity implies meta-charity, it should be common knowledge that all agents are rational. One could no more pretend to be an irrational agent than pretend to be a married bachelor.

At best, drivers could feign being non-agents. Drugs, intense emotion and illness can put human beings into a state that resists interpretation in terms of beliefs and desires. Since only agents deserve punishment, there is a motive to drop out of the moral realm by scuttling the infra-structure for rationality.

People will also act crazy for fascinatingly banal reasons. Consider how social psychologists debunk hypnosis by requesting subjects to pretend they are hypnotized (Wagstaff 1981). With no more motive than a desire to please the experimenter, these obedient fakers drop their trousers, eat onions as if they are apples, and uncomplaining allow pins to pierce their hands. The experimenters never conclude that their bizarre subjects are irrational. The subjects are merely under pressure to meet demands whose weight tends to be under-estimated even by the subjects themselves. In a series of classic studies, social psychologists have demonstrated that people like you and me will go to extraordinary lengths to be punctual and to earn the approbation of strangers. People like you and me will suffer intense pain to avoid embarrassment and lie to avoid being the lone dissenter on a trivial issue. Although we may not want to be strongly motivated by "petty" concerns, this higher order desire does not undermine the rationality of the "crazy" actions performed under the influence of the first order desires.

The involuntariness of rationality cuts both ways. Meta-charity suggests that we cannot choose to be rational. This runs against the grain of many biographies. When one reads Benjamin Franklin's early writings, one gets the impression that personal reversals led Franklin to resolve to become more rational and to interpret others more charitably. These resolutions appear to have made Franklin exceptionally rational and exceptionally charitable. When trying to avert a war between Great Britain and its American colonies, Franklin publicly exhorts both sides to be more rational (rather like a divorce court judge). Privately, Benjamin Franklin fears that the parties will instead irrationally choose to indulge their pride.

Meta-charity implies that diplomats have no basis for fears of irrationality. Indeed, meta-charity implies that I have botched the description of Benjamin Franklin's apprehension; he could not have feared irrational choices because he realized that all agents are rational.

Some people trace part of their recovery from mental illness to their own will-power. John Nash is a recovered paranoid-schizophrenic who won a Nobel Prize in 1994 for a pair of seminal articles in game theory in 1950 (written at the age of twenty two). After struggling up from madness, Nash compared rationality to dieting (Nasar 1998, 353-354). He rejected delusions in the manner of an abstemious fat man turning down sweets. Just as a dieter generalizes his bans to err on the side of safety, John Nash abandoned political thinking. Next went religion. One certainly gets the impression that Nash did not regard himself as rational, that he was instead struggling toward rationality. Since he was *choosing* to censor his thoughts in a campaign to become rational, he seems to count as an agent *prior* to becoming rational.

As one of the seminal thinkers in the mathematical study of rational choice, Nash was drawn to the principle of charity. (Although sometimes he seems to be in the curious position of applying the principle of charity to others but not himself!) This must make Nash's choice to be rational seem paradoxical even to himself. If Nash really chose to be rational, then he was an agent. By charity, he would then already be rational. And then by meta-charity, it would follow that he would recognize the redundant nature of that choice. You cannot choose to acquire what you believe you already possess.

6. Charity is not self-excluding

Sometimes the principle of charity is restricted to a subset of beliefs. If the principle were qualified so that it did not apply to itself, then charity would not imply meta-charity. However, when we examine the standard exclusions, we see that the principle of charity does not have the kind of properties that make a belief a candidate for exclusion.

Despite Davidson's rhetorical question about the earth, many commentators find it quite plausible that the ancients were in massive error about the earth. The intuition is buttressed by causal theories of reference (McGinn 1977). The subject matter of a belief can be fixed by causal relations. In particular, the ancients had beliefs about the earth by virtue of their perceptual contact with it. Their beliefs about the properties of the earth may have played a role in fixing the reference of `earth' but did not supply a synonym for `earth'.

The principle of charity can still be defended by excluding relational beliefs from its domain of application (Vermazen 1982). Notional beliefs are fixed in the way Davidson describes. So Davidson could still maintain that massive error is impossible for general beliefs that make no reference to particular things. The principle of charity, that all agents are rational agents, is just this sort of *general* belief.

The exclusion of relational beliefs can also motivate an exclusion of esoteric beliefs based on authorities. These are the sort of beliefs that we are tempted to put in quotation marks. The physics student believes that "Space is expanding" is true because his teacher says "Space is expanding". The student does not understand how space could expand because he is puzzled as to what space could expand into. But he does not let this conceptual objection stop him from deferring to the experts on space. Surely, top physicists must have anticipated and answered this question. Nor does the student understand what an impacted molar is. Yet he is willing to pay his dentist to cut into his mouth and extract the impacted molar. The student exerts little quality control over the esoteric statements he accepts on authority. Consequently, he is at the mercy of others when it comes to the merit of the statements. Accordingly, he gets little personal credit for

believing $e = mc^2$. He gets little demerit for believing the contradiction a prominent logician asserted. But the inappropriateness of praise or blame is compatible with his beliefs having properties that fit the stereotype of irrationality. If his authority's statements are incoherent, then the content of his echoing belief is also incoherent. If he does not realize that his authorities conflict, then the esoteric statements he believes may be jointly inconsistent. They might be downright meaningless if the communication is garbled.

If the authority is coherent, then his confused followers can be conditionally coherent. They have a disposition to follow the leader, so the authority's beliefs give useful information about followers. If the leader's remarks are incoherent or if the "follower" does not successfully defer to an authority, then the predictive pay-off is so low that belief attribution becomes minimal.

Charity is focused on the narrow content of beliefs. A person who makes rational but unfortunate choices about his authorities can wind up subscribing to intellectually appalling propositions. This "external irrationality" is a common mishap for religious believers. Just as few people do independent thinking about physics and dentistry, few people do independent thinking about religious matters. Religious affiliation is mostly a matter of accepting an authority. Most people are "born into" a religion and so never *choose* their religious authorities. Few people have the means to evaluate religious authorities. Moreover, religions tend to gravitate toward matters in which there is little opportunity for testing. Thus the comparative shopper is unlikely discover any differences that merit a switch from his native religion. Religious affiliation is stable and most conversions are limited to switches based on pragmatic considerations.

When there is little prospect of predictive pay-off people do not bother to interpret what is said. In baseball, a batter does not try to divine the beliefs that underlie the catcher's patter. Religious utterances give especially meager guidance as to how people will behave. An important exception concerns rituals and restrictions on diet, oaths, and military service. Often knowledge of one's neighbor's religion boils down to a short list of operational do's and don'ts (mostly don'ts).

When discussing itineraries or lawn care or gas prices, people are generally attentive and dispassionate about the logical consequences of what they affirm. They readily put what they say into action. The deferential aspects of religious discourse prevent much from being predicted. When religious people act strangely it is often tempting to trace their action to one of their strange religious beliefs. But this kind of resemblance thinking ("Strange effects come from strange causes") cannot explain why the puzzling agent acted at that point of time rather than earlier. Nor can it explain why he acted on that strange "belief" rather than the hundreds of other disturbing sayings that riddle religious texts. And why that agent rather than another member of his congregation? Little wonder that social scientists focus on non-ideological factors when explaining panics, mass suicides and massacres (Kalyvas 1999). They dismiss the journalists' attention to alien religious beliefs as sensationalism.

Alvin Goldman (1992) notes that we lavishly attribute false beliefs to victims of misleading evidence. When police conduct a sting operation, they calculate how their victims will behave by using their victim's evidential base rather than their own. At best, the police maximize their agreement with their victims only with respect to conditionals such as `If I win a valuable prize, I ought to go where it can be collected'.

Goldman's exception to the principle is plausible. But the principle of charity cannot fall into this exempted class. The principle of charity is a general principle. Like all universal generalizations, the principle is a quantified *conditional*. Misleading evidence about whom to subsume under the principle of charity is irrelevant. Any truth can be the premise of an unsound argument. The only misleading evidence that could bear on the principle itself is theoretical evidence. Believers in charity acknowledge that some theorists espouse theories that conflict with charity. But believers in charity must paternalistically attribute charity to these theorists to make sense of the theorist's interpretive success. For instance, if Donald Davidson invites Stephen Stich to give an anti-charity lecture at Berkeley, Davidson will need to explain how Stich managed to navigate from New Jersey to California. He cannot credit Stich's "misleading evidence" about the principle of charity with the power to actually mislead Stich. To interpret a person as misled,

Davidson must rationally reconstruct Stich's error. The rational reconstruction assumes the principle of charity. Consequently, Davidson cannot consistently describe a scenario in which there is misleading evidence against the principle of charity.

The principle of charity needs to be sensitive to perceptual and cognitive asymmetries between people. I ought not attribute my visual beliefs to a blind man. And I ought not attribute my beliefs about the game Kriegspiel to a man who has never heard of the game. The principle of charity also allows me to apply the reasoning in reverse – postulating perceptual and cognitive asymmetries to explain apparent failures of charity. An early suspicion that some men are color blind arose amongst Puritan sects with somber dress codes. Pious Puritans would show up to religious services with gayly colored scarves. Drill sergeants were among the first to suspect dyslexia. A small percentage of soldiers could only tell their left foot from their right foot when the left shoe was marked with straw. Instead of attributing irrationality, we attribute color blindness or an incapacity to discern left from right.

Friends of charity handle asymmetries by wielding the principle in a holistic fashion. Attributing a little falsehood here and there is the best way of maximizing the overall attribution of true beliefs.

Regardless of whether the holistic approach works, we can see that no special measure is needed for the principle of charity itself. Access to the principle of charity is symmetrical. The principle is not esoteric knowledge. It is the sort of thing that is available to each person by virtue of his interpretive competence.

> The concepts of belief, desire, and meaning are common property. The theory that implicitly defines them had better be common property too. It must amount to nothing more than a mass of platitudes of common sense, though these may be reorganized in perspicuous and unfamiliar ways. Esoteric scientific findings that go beyond common sense must be kept out, on pain of changing the subject. (Lewis 1983, 111-112)

7. An ad hoc pocket of irrationality?

Given the failure to find a principled way of preventing the principle of charity from applying to itself, one might just conclude that, as an empirical matter, there just is a pocket of irrationality about the rationality of others. Like it or lump it. Although people are rational, people interpret other people as being irrational. Their mistake is compatible with charity but refutes meta-charity.

Some of this empirical evidence is from the social sciences. There are many studies of "third person effects" in which unfavorable characteristics (cheating, irresponsible drug use, gullibility) are ascribed much more heavily toward others (Davison 1983). For instance, censors deny that they are corrupted by any of the questionable material they examine but ban some of it on the grounds that it will corrupt the masses. One technique of psychological warfare is to distribute propaganda leaflets to the enemy. The evidence suggests that these crude exhortations do not persuade enemy soldiers to surrender or defect or mutiny. However, there is evidence that the leaflets stimulate fear of their persuasiveness. Generals withdraw "vulnerable" units and promote reliability at the expense of efficiency. Last minute "revelations" in a close political campaign do not change many preferences but do alter perceptions of who is likely to win. Undecided voters prefer to vote for a winner. So although the last lob of mud is dismissed by virtually all as irrelevant character assassination, it can still decide the election by our willingness to think many others cannot spot the irrelevance.

The third person effect has reflexive wobbles. After all, the third person effect can itself be the object of third person effects. The third person effect rests on the attribution of an indexical myth which people infer via an egocentric and egotistical fallacy. If I believe that only *others* violate the method of empathy in this way, then I seem to be arbitrarily assigning them just the sort of vice that turns the pistons of the third person effect.

"Grin and bear it!" says the empirical social scientist. Or he might appeal to reflective equilibrium. An account of rationality must consist of the best fit between principles and intuitions. If there is a systematic irreflexivity to interpretive practice, then the theorist must

respect the exception. What seems irregular from one perspective may be uniform from another. Since people's judgments of rationality are constitutive of rationality, the exception is probably the result of an overly simple outlook on rationality.

This attempt to smooth out the irregularity over-estimates the uniformity of the irregularity. People do interpret each other as rational in a variety of circumstances that rivet the attention of economists and game theorists. If you become separated from a friend in Paris, where should you go to re-join him? Most people reach their answer by reasoning about their friend's reasoning. This replication involves the assumption that the lost friend is rational and that he recognizes the rationality of those trying to find him. For instance, one common answer is to meet at the top of the Eiffel Tower (the most salient floor of the most salient building in Paris).

Methods of dividing goods also seem to bring out belief that others believe one to be rational. Think of how people recognize the fairness of requiring the cake-cutter to choose last. They take the perspective of the cake cutter who in turn takes the perspective of earlier cake choosers. The earlier choosers will take a largest available piece.

Irrationality is legendary in the realm of courtship. But social psychologists do not wind up confirming the legend. Many of the apparent "irrationalities" are revealed to be rational effects of intense desires and risky strategies for satisfying those desires. The psychologists portray participants in the mate market as reliable replicators of each other's reasoning. Each participant sees himself or herself through the eyes others and assumes that others are similarly empathic. Each views the other as possessing a realistic appraisal of their value as a romantic partner. Each expects image management. Each expects counter-measures to avoid an over-estimate of the other's package of attractions. The offers and threats constituting courtship are gauged accordingly. In this way romance is unromantically analyzed with bargaining theory. Psychologists chip away at the appearance of irrationality at both the object level *and* the meta-level. Charity and meta-charity creep in together because they are organically related. The rationale for invoking the principle of charity gives meta-charity a free ride. Meta-charity cannot be left behind as an independent after-thought.

Many "irrationalities" are more patiently described as arationality. We prefer explanations that involve temporary suspensions of our agenthood, mental slips, over ones that violate charity. Educationalists engaged in error analysis of arithmetic assume children have a limited budget of attention, memory, and perception. "Stupid" mistakes are analyzed away as mechanical breakdowns. Just as we do not consider a stopped clock as irrationally fixated on a single temporal judgment, we do not consider a boy irrational when he transposes digits in an addition problem. We leave the realm of belief-desire psychology and picture him as having lapsed into a robotic malfunction.

Psychologists frequently force errors by not giving their subjects time or room or perceptual access. At first glance, the overwhelmed subjects appear to be committing stupid mistakes. But the irrationality disappears on close inspection. In the final analysis, these psychologists portray people as doing the best they can given scarce resources. Just as the optical fusion of colors of pointillist picture become invisible on close inspection, the irrationalities of subjects disappear from view when one examines the fine structure of the action.

Daniel Dennett sometimes allows for a little irrationality:

> How rational are we? Recent research in social and cognitive psychology (e.g., Tversky and Kahneman 1974; Nisbett and Ross 1980) suggests we are only minimally rational, appallingly ready to leap to conclusions or be swayed by logically irrelevant features of situations, but his jaundiced view is an illusion engendered by the fact that these psychologists are deliberately trying to produce situations that provoke irrational responses -- inducing pathology in a system by putting strain on it -- and succeeding, being good psychologists. (1987b, 52)

Dennett rounds out his defense by casually observing that judgments of rationality are comparative. The "thinkers" produced by artificial intelligence researchers seem far less rational than children.

Sometimes, Dennett wields responses that preclude irrationality. For instance, Dennett has an amusing discussion of a boy who gives him wrong change for a lemonade purchase (1987a, 84-86). Although the boy has committed an *a priori* error, it is difficult to pin an irrational belief on him. The candidate irrationalities are relevant false propositions about sums and the definitions of `quarter', `nickel', and so forth. But understanding them precludes belief in them. We cannot rationally reconstruct the boy's error. We can only postulate a lapse of agency in which the boy malfunctions like a malfunctioning coin counting machine.

Given that we are maximizing the absolute number of true beliefs or rationality (and are not aiming for a high ratio), surds should be minimized. I applied this theorem controversially to a criticism my wife made of a documentary. Zoo records reveal that half of panda births yield twins. The mother panda chooses to rear only of the twins. Panda experts were trying to rear the rejected pandas in the hope of "doubling the number of pandas reared in zoos". My wife objected: "That only increases the number by 50%, not 100%." After initially congratulating my wife on her astuteness, I had a second thought: "Well, maybe they were restricting the calculation to twin pandas." The Mrs. thought I was bending over backwards to spare the panda researchers.

When in a charitable mood, I try to minimize how many mental slips I attribute. Better to have an expansive believer, unpunctured by holidays from the intentional stance. When in a rarer meta-charitable mood, these rescue missions seem more like an unchivalrous sacrifice of one agent (here my wife) for another agent.

9. A posteriori arguments for charity

A posteriori arguments for charity are generally dismissed as inevitably question-begging (Stein 1996, 114). A critic of charity will balk at a premise such as `All observed agents have been rational'. The critic thinks that it is a platitude that agents are commonly irrational and that this platitude has been richly confirmed by specialists in the psychology of reasoning (Stich 1990, Stein 1996). Only a transcendental argument can undercut this apparent empirical evidence for widespread irrationality.

Nevertheless many proponents of charity appeal to evolution. They reason that since rationality increases reproductive success, natural selection ensures that current human beings are rational. Unless qualified, this mode of argument is incompatible with the thesis that rationality is constitutive of belief (Devitt and Sterelny 1995, 248-249). For it presupposes the possibility of irrational agents. Indeed, the argument seems to admit that irrational agents existed in the past. Natural selection must have selected against the irrational people.

Daniel Dennett (1987a), who favors both the thesis that charity is constitutive of belief and the evolutionary argument, can be consistently interpreted as merely presenting evidence that it is appropriate to adopt the intentional stance toward current human beings. If someone has beliefs, then they must be rational beliefs. Natural selection merely eliminated those individuals who were not really believers. Thus the evolutionary argument concerns an empirical pre-condition for the application of the principle of charity rather than a proof of charity itself.

The same strategy could be used to reconcile other apparent conflicts between the constitutive thesis and empirical arguments. A charitable believer in the design argument for God's existence could consistently argue people are rational because they were designed by a Rational Creator in His likeness. The *a priori* aspect of Dennett's charity is the independently derived principle that all agents are rational agents. The empirical aspect is confined to showing that agents exist.

10. Common knowledge

Since violators of the principle of charity must appear irrational to the subscribers of the principle, the principle of charity indirectly instructs its subscribers to interpret other agents as fellow subscribers to the principle of charity. And if I must view my brethren as fellow subscribers, I must also view my brethren as viewing me as a fellow subscriber. The cycle of principle attribution spirals up indefinitely.

Common knowledge is relativized to populations. If I mistake you as a corpse or a mannequin, then I will not adopt the "intentional stance" toward you. As Dennett stresses, I may

adopt another stance toward you (the design stance or the physical stance). Your behavior may be explicable in the vocabulary of this alternative perspective. But I will not view you as an agent. Belief in the principle of charity must be read *de dicto* (`Mr Charity believes all agents are rational agents') rather than *de re* (`Of each agent, Mr. Charity believes that agent to be a rational agent'). If Miss Deceit fools Mr. Charity into believing she is not an agent, then Mr. Charity will not attribute the principle of charity to her. She will know this, so she will know that Mr. Charity is a rational agent who fails to attribute the principle of charity to another rational agent. But this is not a counterexample to `Charity implies meta-charity' because Mr. Charity's omission would be unwitting. Once it is revealed that Miss Deceit is an agent who is pretending to be a non-agent, Mr. Charity will charitably interpret Miss Deceit. Charity is common knowledge in a community of agents whose agency is common knowledge amongst them. Or at least that is what I say is the commitment of the believer in charity.

The commitment is vexing because the principle of charity does not seem to be common knowledge. Granted, linguists have demonstrated how it is possible to unconsciously follow a syntactic rule that one sincerely (and intelligently) denies following. However, these syntactic rules are only presented as universally *believed*; they are not presented as common knowledge. Common knowledge is the stronger condition in which there is universal belief about the universal belief and in which there is a universal belief about that belief, and so on. It is one thing to be told that you believe what you sincerely deny. It is quite another to be told that you believe everyone (yourself included) believes what you sincerely deny.

Could the principle of charity be merely unconsciously believed by most interpreters? As Ray Jackendoff (1987) notes, attributions of unconscious beliefs are plausible for low level cognition. There is no need for a hunter-gatherer to be aware of raw sensations, nor the principles that are used to convert them into meaningful units. Awareness should be confined to intermediate levels of cognition. But this is just where charity operates. Charity and meta-charity are principles of *interpretation.* Consciousness is for trouble-shooting. Interpretation is applied

to anomalous behavior. Hence, belief in charity and belief in meta-charity seem to be just the sort of beliefs I should be aware of having.

If the principle of charity were common knowledge, then it could not be *learned* from a teacher. The teacher would know the students already know the principle. Yet my impression is that I did learn the principle of charity from a teacher. At the time, the principle struck me as a counterexample to the generalization that nothing practical can be learned from theoreticians. I felt I had been given an into people and proceeded to apply it in daily life. Instead of dismissing dissenting interlocutors as confused or biased or self-deceived, I minimized differences. I became eager to diagnose disputes as merely verbal. I traced what little difference remained to small but critical variations in what we believed and desired. The principle of charity made me a gentleman.

Well, not a perfect gentleman. In the midst of all this rewarding diplomacy, I made a long, unwitting exception for violators of the principle of charity. I did not minimize my differences with *them.* I dismissively criticized those who attributed sloppy thinking to others. I was egged on by articles such as Gerald Massey's "The Fallacy behind Fallacies" and L. Jonathan Cohen's "Whose is the Fallacy?". These authors savor the poetic justice in the accuser being the guilty of the very charge he levels at other. But true believers in charity are in no position to relish the irony of the rubber and glue effect. They cannot ridicule ridiculers – as Peter Winch (1964) does in his aggressive defense of the Azande against Evans-Pritchard's accusations of irrationality. They cannot deflate the deflators. Charity must be equally extended to both the accuser as well as the accused. Charity means being on everyone's side. After all, fallacy mongers often mix it up with each other. The proponent of charity must interpret each party as rational. In this process of rationalization, each fallacy-monger emerges as "consistent, a believer of truths, and a lover of the good (all by our own lights, it goes without saying)" (Davidson 1970, 97). So even their accusations of irrationality must be interpreted as rational. As mentioned earlier, this can be done vacuously by treating the discourse non-cognitively or non-literally.

11. The perils of weakened charity

There are weak variations of the principle of charity that do not have the consequence of common knowledge. If, as Quine suggests (1960, 59) in a moment of moderation, the principle of charity merely forbids the attribution of *blatant* contradictions, then I am free to attribute milder irrationalities. The trouble with these dilute variations is that they cannot perform the roles associated with the principle of charity (overseeing translation, congealing belief-desire explanations, accounting for market efficiency, solving coordination problems, etc.). Quine sometimes writes as if the weak principle were enough to show that we must translate the logical connectives as classical. But any casual dialogue requires an enormous, varied background of inferences. At a minimum I need to interpret my informant as having a conversation with me – even in the inhospitable circumstances of radical interpretation. This alone brings in the kind of calculations to which Paul Grice (1989, 22-40) first drew attention. Dilute versions of charity cannot sustain these calculations. Strong versions of the principle can do the work but only at the price of appropriating for themselves status as common knowledge.

The dilemma arises under all ways of weakening the principle. Strategists recommend charity as an operating assumption because it leaves you prepared for your adversary's strongest response. This minimizes the maximum damage they can inflict while putting you in a good position to exploit an inferior response. The strategists are not predicting that your adversary will choose rationally.

Although this pragmatic attitude toward charity may be useful for strategic purposes, it cannot perform the roles desired by social scientists and philosophers. Nor can they get by with a mere empirical generalization. An empirical generalization presupposes some way to identify the agent's beliefs and desires. If we do not know what their beliefs and desires are, we cannot assess their rationality. Ian Hacking is *criticizing* Donald Davidson when he says that the principle of charity is a rule of thumb "that might, like all common sense, sometimes offer bad advice. "(1975, 149-50) For Hacking is presupposing that the principle of charity is not needed to discover what people believe.

Another weakening uses statistical loopholes. Jonathan Adler (1996, 335) reconciles the principle of charity with the attribution of fallacy by appealing to reference classes. Adler notes that the beliefs logicians pick on are not a representative sample of our beliefs. The vast majority of our beliefs are too trite to merit discussion. Since fallacy attributers share these trite beliefs, they already satisfy Davidson's quota of having almost complete agreement. Even if all their accusations of irrationality were correct, they would still be in massive agreement with those they criticize. Charity always applies to the background, never the foreground.

Adler's compatibilism would stop charity from guiding interpretation. The typical advocate of charity thinks twice before accusing someone of fallacious reasoning. But Adler's thesis implies that the very act of challenging a belief makes it fair game. The principle of charity only protects those beliefs that are not attacked. Adler's conception of charitable interpretation resembles Robert Frost's conception of banking: "A bank is a place where they lend you an umbrella in fair weather and ask for it back when it begins to rain." True, the principle of charity would still protect beliefs from an attack on them as a collective. For the principle still precludes massive error. Thus Adlerian charity may be a resource to the epistemologist in his ancient duel with the skeptic. But this is not the main service for which charity has been employed.

Charity can be diluted in other ways. Sometimes an occasional aberration is permitted. Sometimes rationality becomes merely the central tendency of action. Qualifiers must also be mindful of an inverse relation between the degree of the dilution and the amount of work that can be done by the principle of charity.

12. Extension to the principle of humanity

Arguments for the entailment thesis with respect to charity are readily extended to the principle of humanity. This alternative to charity says that others reason from their perspective as I reason from mine. Hence, I can interpret them by hypothetically adopting their beliefs and desires. According to Richard Grandy (1973), the principle of humanity diverges from the principle of charity when there are asymmetries of information. I see a colleague with a big chalk mark on the

back of his suit. At least at first blush, the principle of charity instructs me to attribute awareness of the chalk mark to my colleague. However, the principle of humanity does not make the implausible attribution because it is sensitive to differences in perceptual access. (Holistic defenders of charity complain that Grandy is applying charity piecemeal rather than systematically. After all, I do need to spare my colleague's belief that chalk marks are easily remedied fashion flaws and easily remedied fashion flaws should be promptly remedied. An interpretation that spares these more central beliefs is more charitable than one that maximizes perceptual similarity.)

Just as charity implies meta-charity, humanity implies meta-humanity. The principle of humanity is reminiscent of the method of empathy: I try to understand you by putting myself in your shoes. So when I try to understand your effort to understand other people, I simulate your interpretive efforts. Since I use the principle of humanity in my interpretations, I thereby impute humanity to you. The principle of humanity is carried along in my meta-interpretation and so becomes embedded in the position I attribute to you. Meta-interpretation breeds meta-humanity.

There may also be extensions to other principles that are related to charity. David K. Henderson (1994) derives charity from the principle that errors be explainable. The question of whether this principle of explicability entails meta-explicability will turn on the details of what counts as an inexplicable error. If someone’s attribution of an inexplicable error would itself count as an inexplicable error, then explicability entails meta-explicability.

The devil is in the details. Nevertheless, a general tendency has been uncovered. Discussions of interpretive principles focus on the simple case of a passive interpretee. The problem has been to pull beliefs and desires from actions as one might extract pearls from oysters. Interpretive principles are then wielded asymmetrically as tools to pry open other minds. But once we consider *active* interpretees who themselves are interpreting others, we see that they also need interpretive principles. Given the symmetry between us and our interpretees, there is pressure to assign the same interpretive principles. Awareness of the symmetry should

itself be symmetrical, so there is further pressure to attribute mutual awareness of the pressure. Consequently, interpretive principles have a general tendency to be common knowledge.

* A version of this paper was read to an audience at York University. I thank members of audience for their comments, especially Henry Jackman. I also thank two anonymous referees.

REFERENCES

Adler, Jonathan (1996) “Charity, Interpretation, Fallacy” Philosophy and Rhetoric 29/4: 329-343.

Alexander, Judd H. (1993) In Defense of Garbage (Westport Connecticut: Prager).

Churchland, Paul (1979). Scientific Realism and Plasticity of Mind, (New York: Cambridge University Press).

Cohen, L. J. (1981) “Can human irrationality be experimentally demonstrated?” Behavioral and Brain Sciences, 4, 317-70. [Includes peer commentaries and response by author.]

Cohen, L. J. (1980) “Whose is the Fallacy?: Rejoinder to Daniel Kahneman and Amos Tversky” Cognition 8: 89-92.

Davidson, Donald (1984a) “On the Very Idea of a Conceptual Scheme” in his Truth and Interpretation (Oxford: Clarendon Press). Originally appearing in Proceedings of the American Philosophical Association (1973/74) XLVII: 5-20.

_____ (1974) “Belief and the Basis of Meaning” Synthese XXVII 3/4: 309-324.

_____(1984b) “Thought and Talk” in his Truth and Interpretation (Oxford: Clarendon Press). Originally appearing in Mind and Language ed. S. Guttenplan (Oxford: Clarendon Press, 1975).

_____ (1976) “Hempel on explaining action” Erkenntnis 10: 239-253.

Davison, W. Phillips (1983) "The Third-Person Effect in Communication" Public Opinion Quarterly 47/1: 1-15.

Dennett, Daniel (1987a) “Making Sense of Ourselves” in his The Intentional Stance (Cambridge, Mass.: MIT Press).

Dennett, Daniel (1987b) "Three Kinds of Intentionality" in his The Intentional Stance (Cambridge, Massachusetts: MIT Press).

Devitt, Michael and Sterelny, Kim (1995) Language and Reality (Cambridge, Mass.: The MIT Press).

Dostoevsky, Fyodor (1918) Notes from the Underground (New York: MacMillan Company).

Goldman, Alvin (1986) Epistemology and Cognition (Cambridge, Mass.: Harvard University Press).

Goldman, Alvin (1992) “Interpretation Psychologized” in his Liaisons (Cambridge, Mass.: The MIT Press) 9-34.

Grandy, Richard (1973) "Reference, Meaning, and Belief" Journal of Philosophy 70 (1973) 442-43.

Grice, Paul. (1989) Studies in the Ways of Words (Harvard University Press).

Hacking, Ian (1975) Why Does Language Matter to Philosophy (Cambridge University Press).

Henderson, David K. (1994) “The Principle of Charity and the Problem of Irrationality (Translation and the Problem of Irrationality)” Readings in the Philosophy of Social Science ed. Michael Martin and Lee C. McIntyre (Cambridge, Mass.: The MIT Press) 323-341.

Hempel, Carl (1962) “Rational Action” Proceedings and Addresses of the American Philosophical Association 1961-1962, vol. XXXV, October, 1962 (Yellow Springs, Ohio: Antioch Press)

Hintikka, Jaakko (1962). Knowledge and Belief Ithaca: Cornell University Press.

Jackendoff, R. (1987) Consciousness and the computational mind (Cambridge, Mass: MIT Press).

Kalyvas, Stathis N. (1999) "Wanton and Senseless? The Logic of Massacres in Algeria" Rationality and Society 11/3: 243-285.

Lewis, David (1983) “Radical Translation” in his Philosophical Papers (New York: Oxford University Press) 108-118.

Linsky, Leonard (1968). `On Interpreting Doxastic Logic'. Journal of Philosophy 65: 500-2.

Massey, Gerald (1975) “Are there any Good Arguments that Bad Arguments are Bad?” Philosophy in Context 4/1: 61-67.

_____ (1981) “The Fallacy behind Fallacies” Midwest Studies in Philosophy: 489-500.

McGinn, Colin (1977) “Charity, Interpretation, and Belief” The Journal of Philosophy LXXIV no. 9: 521-535.

Mehra J. (ed.) (1973) The Physicist's Conception of Nature (Dordrecht: Kluwer).

Nasar, Sylvia (1998) A Beautiful Mind (New York: Simon and Schuster).

Nisbett, Richard, and Ross, Lee (1980) Human Inference: Strategies and Shortcomings (Englewood Cliffs, N. J.: Prentice Hall.

Popper, Karl (1957) The Open Society and Its Enemies (London: Routledge and Kegan Paul).

Sorensen, Roy (1983) "An Essential Reservation Concerning the Efficient Market Hypothesis" The Journal of Portfolio Management 9/4: 29-30.

Stein, Edward (1996) Without Good Reason (Oxford: Clarendon Press).

Stich, Stephen (1990) The Fragmentation of Reason (Cambridge, Mass.: MIT Press).

Tversky, Amos, and Kahneman, Daniel (1974) "Judgment under Uncertainty: Heuristics and Biases" Science 185: 1124-31.

Vermazen, Bruce (1982) “General Beliefs and the Principle of Charity” Philosophical Studies 42: 111-118.

Wagstaff, Graham (1981) Hypnosis, Compliance and Belief (New York: St. Martin's Press).

Wilson, N. L. (1958-9) "Substances without Substrata" Review of Metaphysics 12

Winch, Peter (1964) “Understanding a Primitive Society” American Philosophical Quarterly 1: 307-24.