Formal Ethical Principles

Harry J. Gensler

Abstract
In this paper, Harry Gensler discusses formal ethics, which studies rational patterns in our ethical thinking. He describes four fundamental principles that he calls [r] (a rationality axiom), [e] (ends-means consistency), [p] (prescriptivity) and [u] (universalizability). Gensler also discusses the so-called golden rule ("treat others as you want to be treated") and shows how several versions of this principle can be derived from his axioms. According to Gensler, there are both good and bad versions of the golden rule. One of the good versions can be formulated in the following way: Treat others only as you consent to being treated in the same situation. Gensler shows how this version of the golden rule can be used in our moral thinking and how it can be defended against many common objections. Together the principles discussed in the paper can be used to help us think more rationally about morality and live more consistent lives. The paper brings together several ideas that Gensler has been working on for more than 50 years.

Introduction
"If X does A to you, then do A to X" is a formal ethical principle. A *formal ethical principle*, as I use the term, is an imperative or ethical principle expressed or expressible using only variables (for things like persons, actions, and propositions) and abstract logical or quasi-logical notions (like logical terms; terms for general psychological attitudes, like *believe*, *desire*, and *act*; and other fairly abstract notions, like *ought* and *ends-means*).[1] Here are examples to clarify the idea:

- FORMAL: "Treat others as they treat you." This is a *formal ethical principle*, since we can express it as "If X does A to you, then do A to X," where X is a person variable and A is an action

[1] We could express this definition more precisely by listing *which variables* and *which abstract logical or quasi-logical notions* are allowed – or by embedding this list into the wff (well-formed formula) definition of a formal system that symbolizes and systematizes formal ethical principles (as in chapters 12 to 14 of my *Introduction to Logic*, see next footnote). This current paper will be less technical.

variable. (I'll consider "you" and "I" in my examples to be variables, addressed to any agent.)

- NOT FORMAL: "Don't hit others." Here *hitting* is a concrete kind of action and can't be replaced by a general action variable; so this isn't a *formal* ethical principle.

Formal ethics, somewhat patterned after *formal logic*, is the study of formal ethical principles, like my first example above. This paper is an introduction to formal ethics.[2]

Not all formal ethical principles are good principles. "Treat others as they treat you" can lead to endless revenge cycles (I hurt you, so you hurt me, so I hurt you again, … – and hurting multiplies); so this is a *bad* formal principle. Formal ethics tries to separate *good* formal ethical principles from *bad* ones. As with formal logic (in testing for validity or invalidity), we can usually appeal to clear counterexamples or contradictions against bad formulas.

Many common ethical sayings can be expressed as *formal ethical principles*. Here are five examples, expressed first in regular English and then with variables:

1. Be consistent in your beliefs. If A and B are logically inconsistent, then if you believe A then don't believe B.
2. If you want to achieve a goal, then take the necessary means. If you have a goal G and believe that you need to do means M to fulfill goal G, then do means M.
3. Follow your conscience. If you believe that you ought to do A, then act to do A.
4. Evaluate similar cases similarly. If you believe that X in situation S ought to do A, then believe that anyone else in situation S ought to do A.
5. Treat others as you want to be treated. If you want X to do A to you, then do A to X.

These *formal ethical principles* try to express common and important ideas. But each one is formulated poorly and quickly leads to absurdities. I'll try to give better formulations. And I'll try to show how formal ethical principles

[2] See my *Formal Ethics* (New York: Routledge, 1996). These further books of mine are also useful to consult (and I'll refer to them just by title): *Ethics and the Golden Rule* (New York: Routledge, 2013), *Ethics and Religion* (New York: Cambridge, 2016), *Introduction to Logic*, 3rd ed (New York: Routledge, 2017), and *Ethics: A Contemporary Introduction*, 3rd ed (New York: Routledge, 2018).

can give useful tools for reasoning about ethics and for criticizing bad ethical views (like racism).

I'll also try to *systematize* formal ethical principles. The good forms, I contend, prescribe some sort of consistency. And so we'll talk about consistency in beliefs (example 1), consistency in will (examples 2 and 3), impartiality (example 4, also a kind of consistency), and the golden rule (example 5, also a kind of consistency). All of these rest on a broader consistency norm: *We ought to be consistent in thought and action*.

Here I'll stay neutral on foundational questions. You could accept my proposed principles as self-evident truths, divine commands, social conventions, expressions of feeling, what we'd desire if we were ideal ethical thinkers, universalizable prescriptions, or what we need to do to satisfy an ideal of moral consistency. My goal is to explain and defend useful formal tools of ethical reasoning that can supplement and enhance practically any approach to ethics – and can help us to think more clearly and deeply about ethical issues.[3]

---

[3] Ordinary language uses "formal" in many ways. We speak of "formal dress" or "formal agreements"; here "formal" means "following official standards strictly" and contrasts with being loose or casual. Or we speak of the "A-B-B-A form" of a song; here "form" is about structure or arrangement, and is opposed to matter or content; this is closer to my meaning. Even closer is a written "form," a document with blanks to be filled in. But my notion of "form" borrows more from logic: a *formal ethical principle* is an imperative or ethical principle expressed or expressible using only variables and abstract logical or quasi-logical notions.

I've learned much about the formal aspects of morality from Immanuel Kant (especially his *Groundwork of the Metaphysics of Morals*, trans. H.J. Paton [New York: Harper & Row, 1964, first published in 1785]) and Richard Hare (especially his *Freedom and Reason* [Oxford: Clarendon Press, 1963]). But their projects differ from mine; they propose a comprehensive view of the nature of morality, while I keep neutral on this (at least in this paper). Kant and Hare sometimes use "formal" in senses different from mine (even though many of their principles are also "formal" in my sense). Kant (*Critique of Pure Reason*, trans. N.K. Smith [New York: St. Martin's Press, 1965, first published in 1781 and 1787], p. 387) calls a principle "formal" if it deals with objects of every sort; he says that ethical principles aren't formal in this sense, since they must deal with actions. Later in this book he calls a practical principle "formal" if it abstracts from all subjective ends (p. 427); such imperatives have the "form of universality" (p. 431). Hare (*Moral Thinking* [Oxford: Clarendon Press, 1981], pp. 4–5 and 62–4) calls a claim about morality "formal" if it has to do with the meaning or logical properties of the moral words. None of these is what I mean by "formal."

Some thinkers use "formal ethical principle" more broadly, to cover general methodological principles like "Be factually informed" and "Develop your imagination." I sometimes call such

A. Consistency in Beliefs

Consistency in beliefs demands that we not accept logically incompatible beliefs – and that we not accept a belief without also accepting its logical consequences. Suppose that Ima Relativist begins her essay by contradicting herself:

> Since morality is relative, *no duties bind universally*. What's right in one culture is wrong in another. Universal duties are a myth. So *everyone ought to be tolerant of others*.

Ima's first italicized statement ("No duties bind universally") is logically incompatible with her last ("Everyone ought to be tolerant of others"). Ima is confused and inconsistent. She violates the following *consistency imperative*:

> Don't combine these:
>
> - I believe "No duties bind universally."
> - I believe "Everyone ought to be tolerant of others."

Consistency requires that Ima give up one belief or the other; but it doesn't say which to give up. *Consistency norms forbid inconsistent combinations but don't tell us specifically what to believe or what to do*. So consistency norms, while giving us some rational guidance, keep us free to form our own beliefs.

Our consistency imperative here is based on a *formal ethical principle*, which we can express as a *don't-combine imperative* (here "i" is for "inconsistent"):[4]

> (i) If A and B are logically inconsistent, then *don't combine* believing A and believing B.

Why shouldn't we instead express the formal ethical principle as an *if-then imperative*?

> If A and B are logically inconsistent,
> then *if* you believe A *then don't believe* B.[5]

---

principles "semiformal" to distinguish them from ones that are "formal" in my variable-constant sense. In practice, formal and semiformal principles need to work together.

[4] A *consistency imperative* may contain concrete ideas (like "enslave" or "tolerate"), but it must follow from a *formal ethical principle* with no such concrete ideas; this one follows from "If A is logically inconsistent with B, then don't combine believing A and believing B." In formulating consistency imperatives and formal ethical principles, I prefer the brief "Don't combine …" form; but I'd also assert the longer forms, like "You ought not to combine …" or "It's bad to combine these inconsistent things …"

[5] In the formalized version of my theory (see *Introduction to Logic*, ch. 13), (i) is symbolized as "$(\sim\Diamond(A \bullet B) \supset \sim(\underline{u}{:}A \bullet \underline{u}{:}B))$" and is provable as a theorem, while the bad if-then imperative is

What's the difference? The good *don't-combine imperative* (i) tells Ima (who believes each of two logically incompatible ideas) *not to combine* the two; in other words, Ima needs to give up at least one of her two ideas (but it doesn't tell Ima which to give up). The bad *if-then imperative* tells Ima to give up her second belief (since it's incompatible with her first belief) – and to give up her first belief (since it's incompatible with her second belief); so the bad *if-then imperative* tells her to give up *both* beliefs. But consistency shouldn't tell her to give up *both* of the two incompatible beliefs – since maybe one is fine while the other isn't. Consistency should just tell her, "The things you accept don't fit together – so you must change something."[6]

Use don't-combine imperatives, not if-then imperatives. This is an important point about *consistency principles* and *formal ethical principles*; we'll see it again and again, including about the golden rule.

Abraham Lincoln often used consistency against slavery.[7] Suppose that you think that having higher intelligence gives one a right to enslave another. Do you then think that anyone with higher intelligence than you has a right to enslave you? Very few would believe that.

Don't combine these:

- I believe "Anyone with higher intelligence has a right to enslave anyone with lower intelligence."
- I don't believe "If X has higher intelligence than me, then X has a right to enslave me."

Suppose that your principle (which permits you to enslave another) clashes with your concrete judgment (which forbids others to enslave you). Then consistency requires that you change either your principle or your concrete judgment – but it doesn't tell you which to do. Which should you change? Well, it depends; maybe your principle is flawed or maybe your

---

symbolized as "(∼◊(A • B) ⊃ (u:A ⊃ ∼u:B))" and isn't provable. The difference isn't due to the "∼(… • …)" to "(… ⊃ ∼…)" shift (both are equivalent) but to the difference in underlining (which in my system indicates imperative form, see *Introduction to Logic*, ch. 12).

[6] Another problem is that the bad if-then imperative tells people who believe a self-contradiction to believe nothing. Suppose that Ima believes "(P • ∼P)." Following classical logic, "(P • ∼P)" is inconsistent with Q (for any Q). So then, since Ima believes "(P • ∼P)," the bad if-then imperative tells her not believe Q (for any Q); thus it tells her to believe nothing.

[7] See *The Collected Works of Abraham Lincoln*, 8 vols., ed. R. Basler (New Brunswick, N.J.: Rutgers, 1953), 2:223–4. It's wrong to say that we can't dispute ethical first principles; we often dispute them by appealing to consistency.

concrete judgment is flawed; you'll have to think things through more deeply (and maybe apply the golden rule or other items that we'll talk about later).

Our consistency imperative here is based on this *don't-combine imperative* (here "e" is for "entails"):[8]

(e) If A logically entails B, then *don't combine* believing A and not believing B.

Again, it would be wrong to express this as an *if-then imperative* (as either of these):

If A logically entails B,
then *if* you believe A *then believe* B.

If A logically entails B,
then *if* you don't believe B *then don't believe* A.[9]

Suppose that your principle clashes with your concrete judgment; then the first one here tells you to *always change your concrete judgment* (since your principle clashes with it) – while the second one tells you to *always change your principle* (since it clashes with your concrete judgment). But again, what you should do depends on the situation; maybe your principle is flawed (and you need to change it) or maybe your concrete judgment is flawed (and you need to change it). Consistency only prescribes that we have a harmony among our beliefs (e.g., between our principles and our concrete judgments) – but it doesn't tell us specifically what to change to achieve this harmony.[10]

Here's a related example about racism. Suppose that Ima Racist tells us, "We ought to treat blacks poorly – because they're inferior." While we could easily attack his factual premise, I'm more interested in his implicit ethical principle: "All who are inferior ought to be treated poorly." To criticize this,

---

[8] Principles (i) and (e) (for "inconsistent" and "entails") are preliminary formulas that will be expanded later; the final versions will be combined into the two-part [r] principle. None of these principles occurs as such in the formalized version of my theory (see *Introduction to Logic*, ch. 13), which instead uses possible-worlds rules that give the same results but are easier to use in formal proofs.

[9] In the formalized version of my theory (see *Introduction to Logic*, ch. 13), (e) would be symbolized as "(□(A ⊃ B) ⊃ ∼(u:A • ∼u:B))" and is provable as a theorem, while the two bad if-then imperatives are symbolized as "(□(A ⊃ B) ⊃ (u:A ⊃ u:B))" and "(□(A ⊃ B) ⊃ (∼u:B ⊃ ∼u:A))" and aren't provable.

[10] Another problem is that the first bad if-then imperative tells people who believe a self-contradiction to believe everything. Suppose that Ima believes "(P • ∼P)." Following classical logic, "(P • ∼P)" logically entails Q (for any Q). So then, since Ima believes "(P • ∼P)," the first bad if-then imperative tells her to believe Q (for any Q); thus it tells her to believe every proposition Q (including the contradictory of every Q).

we must first clarify its meaning. Suppose that Ima by "inferior" means "of IQ less than 80"; then his principle means:

P All who have an IQ of less than 80 ought to be treated poorly.

But every race clearly has some members of IQ less than 80 and some of IQ greater than 80.[11] So we should remind Ima that his principle P has this consequence about whites:

C All *whites* who have an IQ of less than 80 ought to be treated poorly.

Ima's principle commits him to thinking that he ought to treat many whites poorly (as he treats blacks). As a racist, he won't accept this.[12] But then he's accepting a principle and yet refusing to accept its logical consequences. His views are incoherent. To bring his beliefs into harmony with each other, he must either give up principle P or else come to accept consequence C. One beauty of the appeal to consistency is that it doesn't presume material ethical premises (that the other party may reject) but just points out problems in the other person's belief system.

Consistency norms forbid inconsistent combinations. They don't say to prove all our beliefs (which is impossible), to shun emotions, or to never change our minds. Consistency norms can be defended in various ways – perhaps as self-evident truths, divine commands, or social conventions. They're part of normative ethics (saying that we *ought* to be consistent) and also metaethics (giving conditions needed to be *rational* in our moral beliefs).

Consistency norms aren't exceptionless. They don't apply, for example, if we psychologically can't be consistent, if the logical connections are too complex for us to grasp, or if some stronger duty interferes (perhaps Dr. Evil will destroy the world unless we're inconsistent). All our consistency norms are subject to implicit qualifications. As we deliberate about alternatives, these qualifications usually aren't very important.[13]

---

[11] Ima won't be able to find any other meaning of "inferior" that cleanly divides the races.

[12] If Ima Racist accepts this, then he gives up racism in favor of a kind of elitism: that people of whatever race or group ought to be treated poorly if their IQ drops below a certain level.

[13] *Formal Ethics* §2.3 argues that our formal ethical principles should be qualified by some such phrase as this:

QF You're a person who is (or should be) aware of the logical relationships involved, you have (or should have) at least some interest in the inferred belief (if there is one), you're able to act in the manner specified, and there are no offsetting reasons why you shouldn't do so.

Consistency doesn't guarantee truth (since it's possible to be consistent but wrong), but it often points us toward the truth. Inconsistencies can lead us to see that a person is lying. And finding inconsistencies in a view can help us to discover a better view.[14]

B. Consistency in Will

There's also *inconsistency in will*. Here's an example:

Don't combine these:

- I resolve to eat nothing.
- I eat this granola bar.

If I combine both parts, then I'm inconsistent and need to change something; consistency alone doesn't tell me specifically what to do, since that depends on the situation. If my medical procedure requires fasting, then I maybe should resolve and not eat. But if my fasting is an unhealthy way to diet, then I maybe should eat and not resolve.

Here are further examples of inconsistency in will. I make a firm resolution (to run every day), but then act against it (I put it off and don't do it). I have all-things-considered desires that I know are incompatible (e.g., to

---

We normally satisfy QF when we appeal to consistency. We make the person aware of the inconsistency, the person is able to avoid it, and Dr. Evil isn't threatening to destroy the world unless you're inconsistent. In practice, the QF proviso isn't too important. Since this proviso suggests a disanalogy between formal ethics and formal logic, I note that quantificational logic as applied to English sentences also needs to be qualified by some such phrase as this:

> QL Each term has a constant meaning and reference (including temporal qualifications) in all its occurrences, each statement is clear enough to be definitely true or false, each predicate produces definite truths or falsehoods for all the entities in the universe of discourse, the statements aren't illogical idioms [like "I don't know nothing," which in English doesn't follow logic's rule that "∼∼P" is equivalent to "P"], at least one entity exists in the domain of objects, and the individual constants refer to existing entities.

So formal ethics and formal logic both need qualifiers.

[14] *Formal Ethics* §§2.4–7 discusses some further issues, like: "What is belief?" "Does belief come in degrees?" "Is consistent belief *conjunctive* – so we shouldn't combine believing P_1 and believing P_2 … and believing P_n and not believing (P_1 and P_2 … and P_n)?" "Instead of using consistency imperatives, could we express similar ideas in terms of the *virtue* of consistency (*being a consistent person*)?" "Can we *choose* what to believe?" "Are consistency duties *moral* duties or some other kind of duty?" "How do we respond to someone who denies the law of noncontradiction (as in paraconsistent logic)?" "Since there are seldom conclusive arguments in ethics, how can logic be relevant to ethics?"

become a doctor and to party all the time). As a politician, I endorse incompatible proposals. Or I violate ends-means consistency or conscientiousness.

All our *consistency in beliefs* principles in the last section can be derived from that section's principles (i) and (e) (for "inconsistent" and "entails"):

(i) If A and B are logically inconsistent, then don't combine believing A and believing B.

(e) If A logically entails B, then don't combine believing A and not believing B.

We'll now broaden the wording of these so that they also cover *consistency in will*:

(i) If A and B are logically inconsistent, then don't combine *accepting* A and *accepting* B. *(Here A and B can represent indicatives or imperatives or a mix of these.)*

(e) If A logically entails B, then don't combine *accepting* A and *not accepting* B. *(Here A and B can represent indicatives or imperatives or a mix of these.)*

Differences here are italicized. We'll now treat "believing" as *accepting an indicative* – and we'll treat "willing" as *accepting an imperative*:

- You *believe* that A = You accept (endorse, assent to, say in your heart) "A is true."
- You *will* that act A be done = You accept (endorse, assent to, say in your heart) "Let act A be done."

We'll often use terms more specific than "will" – like "act," "resolve to act," or "want." Which of these fits depends on whether the imperative is present or future, and whether it applies to oneself or to another. Here are some examples:

- To accept a present tense imperative addressed to yourself is *to act to do the thing* (to act with the intention of doing it). So to accept "Eat this" (addressed to yourself) is *to act to eat this*.
- To accept a future tense imperative addressed to yourself is *to intend to do the thing* (to be resolved to do it). So to accept "Eat nothing" (addressed to yourself) is *to intend to eat nothing*.
- To accept a present or future tense imperative addressed to someone else is *to want the person to do the thing* (to desire that the person do it). So to accept "Let everyone be kind to everyone

else" is *to want everyone to be kind to everyone else*.[15]

- To accept a past tense imperative is *to wish that it had been done or regret that it was done*. So to accept "Would that I hadn't done that" is *to regret that you did it (or wish that you hadn't done it)*.

My examples stretch the ordinary language use of "imperative" and "accept an imperative." Here I'm giving a theoretical construct to help us to systematize the notion of *consistency of will*. If you like, you can imagine acts of willing (acting, resolving, desiring, and so forth) as being accompanied by inner speech using imperatives, with the consistency of the willing being gauged by the consistency of the imperatives. Formal ethics needn't in principle use this analysis of willing, as *accepting an imperative*; but this analysis makes it easier to develop formal ethics as a unified system with a few basic axioms – rather than as a mixed assortment of various principles.

Given this structure, we can derive our consistency imperative about eating the granola bar:

1 Imperatives "Eat nothing" and "Eat this granola bar" are logically inconsistent. (This is clear and can be shown easily using imperative logic.[16])

2 If imperatives "Eat nothing" and "Eat this granola bar" are logically inconsistent, then don't combine accepting "Eat nothing" and accepting "Eat this granola bar." (This is an instance of our new principle (i).)

∴ Don't combine accepting "Eat nothing" and accepting "Eat this granola bar." (From the two premises.)

∴ Don't combine resolving to eat nothing and acting to eat this. (This substitutes equivalents about what *accepting an imperative* involves.)

C. Ends-means Consistency

Ends-means consistency requires that we keep our means in harmony with our ends. We violate this if we (1) have an end, (2) believe that to fulfill this we need to carry out certain means, and (3) don't carry out the means.

---

[15] "Desire" and "want" can have a *prima facie* sense ("I have *some* desire to do A") or an all-things-considered sense ("*All things considered*, I desire to do A"). Here I intend the latter.

[16] See *Introduction to Logic* ch. 12, which would symbolize these two imperatives as "(x)~E$\underline{u}$x" and "E$\underline{u}$t." The corresponding consistency imperative would be symbolized as "~($\underline{u}$:(x)~E$\underline{u}$x • $\underline{u}$:E$\underline{u}$t)" and can be proven very simply using ch. 13.

Suppose Maria has the all-things-considered goal to become a doctor. She realizes that, to do this, she needs to study hard and get good grades; but she doesn't act accordingly. Ends-means consistency forbids a combination:

Don't combine these:

- I have the goal to become a doctor.
- I believe "Achieving this requires that I study hard and get good grades."
- I don't study hard and get good grades.

Since Maria's goals, beliefs, and actions don't fit together, she must change something. Maybe her doctor-goal is unrealistic and should be rejected; or maybe she just needs to carry out the means. Consistency doesn't say what to change.

How can we incorporate ends-means consistency into our system? We again have to broaden principles (i) and (e) (for "inconsistent" and "entails") from their previous formulation:

(i) If A and B are logically inconsistent, then don't combine *accepting* A and *accepting* B. *(Here A and B can represent indicatives or imperatives or a mix of these.)*

(e) If A logically entails B, then don't combine *accepting* A and *not accepting* B. *(Here A and B can represent indicatives or imperatives or a mix of these.)*

The final versions of these will be combined into a two-part "rationality" axiom [r] (here the capital letters can represent indicatives or imperatives or permissives or a mix of these – and [e], [p], and [u] are axioms that will be introduced later):

[r] If a set of one or more premises P_1, P_2, ..., P_n – either by themselves or with [e], [p], and/or [u] – lets us derive a self-contradiction, then don't combine accepting each premise.

If a set of one or more premises P_1, P_2, ..., P_n – either by themselves or with [e], [p], and/or [u] – lets us derive a conclusion, then don't combine accepting each premise and not accepting this conclusion.

These are two main changes:

- We now can derive consistency imperatives that involve more than two elements (whereas (i) and (e) are limited to just two elements). We need this, because ends-means consistency imperatives involve three elements (goal, belief, and action).
- Three further axioms ([e], [p], [u]) can be used to derive the self-

contradiction or conclusion. The four axioms in brackets ([r], [e], [p], [u]) will suffice to derive all of the correct consistency imperatives and formal ethical principles discussed in this paper.[17]

Our four axioms together tell us to think and live consistently with logic and three further principles (about ends and means, keeping our moral beliefs in harmony with our lives, and making similar evaluations about similar cases). As before, assume an implicit QF qualifier with each instance of [r] (see the section A footnotes).

We'll now add axiom [e] (for "ends-means") to provide for ends-means consistency:

[e] From "Do A" and "Doing B is a necessary means to doing A," derive "Do B."

I leave open whether [e] is a strict logical entailment (as I think it is) or whether it's not this but rather a principle that we need for other reasons (perhaps because it's socially accepted, commanded by God, or needed to avoid frustrating our goals). I leave this open because I want my *formal ethics* framework to appeal to a wide range of theoretical views.[18]

Given this new framework, we can easily derive ends-means consistency imperatives:

1 Do A. (Premise; here we could substitute a goal, like "Become a doctor.")

---

[17] My system here is based on four axioms, as in my *Formal Ethics* book. Here "axiom" isn't opposed to "inference rule," since most of my four axioms are worded like inference rules.

[18] My *Introduction to Logic* doesn't add this principle to its imperative logic (since "doing B is a necessary means to doing A" would likely require adding causal necessity to my system, which was getting too complicated anyway). Instead, for examples about ends-means consistency, I just add an ad hoc premise like the one below. Here's an example from the problems that students are to work out (p. 302).

"Attain this end" entails "If taking this means is needed to attain this end, then take this means."

∴ Don't combine (1) wanting to attain this end and (2) believing that taking this means is needed to attain this end and (3) not acting to take this means.

This is symbolized as "$\Box(\underline{E} \supset (N \supset \underline{M})) \therefore \sim((\underline{u}{:}\underline{E} \bullet \underline{u}{:}N) \bullet \sim\underline{u}{:}\underline{M})$" and its easy proof (p. 406) takes 13 steps. Here the premise serves instead of axiom [e], which this current paper uses.

Formal Ethics §8.4 suggests symbolizing "Your doing means M is causally necessary for you to do end E" using the symbol "[c]" for causal necessity, as "[c](~M ⊃ ~E)." (More properly, we'd use a small "c" inside a box.)

2 Doing B is a necessary means to doing A. (Premise; here we could substitute a belief about ends and means, like "Studying is a necessary means to becoming a doctor.")

∴ Do B. (From the last two lines, using axiom [e].)

∴ Don't combine accepting "Do A" (goal), accepting "Doing B is a necessary means to doing A," and not accepting "Do B" (means). (From lines 1 to 3 using [r].)

∴ Don't combine intending to do A (goal), believing that doing B is a necessary means to doing A, and not acting to do B. (This substitutes equivalents.)

Note the *don't-combine* form of the derived conclusion.

Why shouldn't we instead express ends-means consistency as an *if-then imperative*?

If you have the goal to do A and
you believe that doing B is a necessary
means to doing A, *then do B.*

This formula has lots of problems, including:

- You may have evil goals; then this formula can tell you to do evil things.
- You may have two goals (maybe the goal to become a doctor and the goal to party all the time) that conflict, given your beliefs; then this formula can tell you to do contradictory things.
- You may have a single goal (maybe to do whatever Mom or Dad tell you) and it may still lead to conflicts, given your beliefs (maybe you believe that Mom tells you to take out the garbage now and that Dad tells you *not* to take out the garbage now); then again, this formula can tell you to do contradictory things.[19]

Having inconsistent goals is a big problem in life. When our goals conflict, we need to somehow qualify them so that they don't conflict (so perhaps you keep the goal to become a doctor but qualify the goal to party all the time by adding the words "after I've done enough homework"). Our *don't-combine* approach would forbid combinations of goals, beliefs, and actions that conflict; in case there's a conflict, we need to do some soul

[19] Some goals are self-contradictory in themselves (without appealing to any further facts or beliefs). Maybe I have the goal to be greater than everyone in the world (including myself); this is impossible to fulfill. Our system would forbid this goal. (Thus when I say that my system forbids combinations, but not specific items, there's an exception here; my system can forbid specific items that are self-contradictory.)

searching and then change something. Here's how to derive a *don't-combine* result for the second case above:

1 Become a doctor. (Premise; your goal.)
2 Party all the time. (Premise; also your goal.)
3 Studying much is a necessary means to becoming a doctor. (Premise; your belief.)
4 Not studying much is a necessary means to partying all the time. (Premise; also your belief.)
∴ Study much. (From 1 and 3, using axiom [e].)
∴ Don't study much. (From 2 and 4, using axiom [e], contradicting the previous line.)
∴ Don't combine accepting "Become a doctor" (goal), accepting "Party all the time" (goal), accepting "Studying much is a necessary means to becoming a doctor," and accepting "Not studying much is a necessary means to partying all the time." (This follows using [r] and the fact that lines 1-4 with [e] leads to a contradiction.)
∴ Don't combine intending to become a doctor (goal), intending to party all the time (goal), believing "Studying much is a necessary means to becoming a doctor," and believing "Not studying much is a necessary means to partying all the time." (This substitutes equivalents.)

Ends-means inconsistency, like belief inconsistency, is common. We often do what's easy or immediately satisfying, instead of what's needed to fulfill deeper goals. While Aristotle defined "humans" as "rational animals," we're imperfectly rational, and our rational and animal dimensions can fight each other.

Some goals are evil. Our ends-means norms don't command evil means to promote an evil goal; instead, they just forbid inconsistent combinations, like:

Don't combine these:

- I have the goal to take maximal revenge against Al.
- I believe "Achieving this requires slashing Al's tires."
- I don't slash Al's tires.

You shouldn't combine these, because (1) the three are inconsistent, and (2) you shouldn't have this goal (and so you shouldn't combine this goal with other things). Consistency, again, forbids a combination; it doesn't say that, if you in fact happen to have this goal and this belief, then you ought to do the

act. In addition, ends and means need to satisfy other consistency principles. Here the ends and means likely violate the golden rule (given that I'm not willing that my tires be slashed in similar circumstances) and clash with my other goals (to avoid a mutually destructive revenge-war).

In the movie *Erin Brockovich*, a power company in order to maximize profits (goal) released toxic water into the ground (means) without cleaning it up; they knew the toxic water would cause cancer to many – and they weren't willing that this be done to them in the same situation. After a law clerk investigated, the company had to pay a fine and clean things up. Their unrestricted goal (to maximize profits, even by means that harm others) and the means (releasing toxic water) fail the golden-rule test. To be fully consistent, our goals and actions need to satisfy the golden rule; a morally responsible company would do this.

Immanuel Kant's formula[20] – "Treat humanity, never simply as a means, but always at the same time as an end" – points to a difference between how we should treat things and how we should treat people. We can use a hammer as a *mere means* to promote our goals; but it's wrong to use others as a *mere means*. It can be fine to use a person as a means, as I use a dentist in order to get healthy teeth. But, at the same time, I must treat persons as persons (and not as things) and take into account what happens to them, as I want others to do toward me; I must treat others only as I'm willing to have myself treated in the same situation. It's evil to treat persons as a *mere means* to promote my goals.

Ends and means are important to practical reason and human life. We have many goals – including food, shelter, health, companionship, and meaningfulness. Practical reason has us try to understand our goals, make them consistent with each other, investigate how to satisfy them, imagine various end results, satisfy ends-means consistency, and reject ends or means that lead us to violate golden-rule consistency.

D. Conscientiousness

Conscientiousness requires that we keep our actions, resolutions, and desires in harmony with our moral beliefs. We violate this if our moral beliefs clash with how we live and want others to live. Suppose I accept pacifism: "One ought never to kill a human being for any reason." If I'm conscientious, then (1) I never intentionally kill a human being, (2) I resolve not to kill for any

[20] *Groundwork of the Metaphysics of Morals*, trans. H.J. Paton (New York: Harper & Row, 1964), p. 96.

reason (even to protect myself or my family), and (3) I don't want others to kill for any reason. Similar requirements cover beliefs about what is "all right" ("permissible"). If I'm conscientious, then I won't believe that something is all right without consenting to (approving of) the idea of it being done; and I won't do something without believing that it would be all right for me to do it.

Conscientiousness says "Avoid inconsistencies between your moral beliefs and how you live and want others to live." Here's an example:

Don't combine these:

- I believe "I ought to do A now."
- I don't act to do A now.

When we combine these, our moral belief clashes with our actions – and consistency requires that we change one or the other. Consistency doesn't say "Our conscience is always right"; if our moral belief clashes with our actions, it could be that our actions are fine but our moral belief is wrong.

To derive conscientiousness imperatives, we'll add axiom [p] (for "prescriptivity"[21]):

[p] From an ought-judgment, derive the corresponding imperative (like "Do it" from "You ought to do it").
From a wrong-judgment, derive the corresponding negative imperative (like "Don't do it" from "It's wrong for you to do it").
From an all-right-judgment, derive the corresponding permissive (like "You may do it" from "It's all right for you to do it").[22]

I leave open whether [p] is a strict logical entailment (as I think it is) or whether it's not this but rather a principle that we need for other reasons (perhaps keeping a harmony between our moral beliefs and how we live is a duty based on social conventions, divine commands, self-evident truths, or something else). Again, I leave this open because I want my *formal ethics* framework to appeal to a wide range of theoretical views.

---

[21] I borrow this term from Richard Hare, who saw *prescriptivity* and *universalizability* as logical features of ought-judgments (and expressing logical entailments). My [p] and [u] are weaker, since they don't claim logical entailments.

[22] "You *may* do it" here is a *permissive*, a weaker member of the imperative family (and not just another way to express "all right"). Accepting "Act A may be done" expresses that you *consent* to it being done (or are *willing* that it be done), although you don't necessarily *want* it to be done (since it may not be your first choice). We can consistently consent both to the act and to its omission – saying "You may do A and you may omit doing A." See *Formal Ethics* §3.1 and *Introduction to Logic* §§14.4–14.6.

In any case, [p] is to be applied only to all-things-considered judgments that express your own evaluation – so "You ought (all things considered, this is my evaluation) to do A" commits you to accepting "Do A"). In contrast, it's perfectly fine to accept "You ought (other things equal) to keep this promise, but (because there are extenuating circumstances) don't keep the promise" – or "You ought (according to department policy) to do this, but please don't (because department policy is really stupid)."

Given [p], we can derive the conscientiousness imperative at the beginning of this section:

1 I ought to do A now. (Premise.)
∴ Do A now. (From premise 1 using axiom [p].)
∴ Don't combine accepting "I ought to do A now" and not accepting "Do A now." (From lines 1 and 2 using [r].)
∴ Don't combine believing "I ought to do A now" and not acting to do A. (This substitutes equivalents.)

Note the *don't-combine* form of the derived conclusion.

Why shouldn't we instead express this conscientiousness idea as an *if-then imperative*?

If you believe "I ought to do A,"
then do A.

This formula has problems, including:

- You may have a deranged moral belief (e.g., you believe that you ought to commit mass murder); then this formula can tell you to do evil things.
- You may have a self-contradictory belief (e.g., you believe that you ought to do A and also ought not to do A); then this formula can tell you to do self-contradictory things.

As before, *use don't-combine imperatives, not if-then imperatives.*

Here's a consistency analogue of "Practice what you preach":

Don't combine these:

- I believe "Everyone ought to do A."
- I don't act to do A myself.

This doesn't assume that our norms are correct (so if we preach universal hatred then we ought to hate). Instead, it forbids inconsistencies between our norms and our actions; if these clash, then something is wrong (maybe our norms).

We can use consistency to criticize basic norms, even ones that seem self-evident. Suppose that my culture taught me to enjoy beating up short people

and to accept *shortism*: "*All short people ought to be beat up, just because they're short*." To accept shortism consistently, I have to *believe* that if I were short then I ought to be beat up – and *desire* that if I were short then I be beat up. Consistency forbids this combination:

Don't combine these:

- I believe "All short people ought to be beat up, just because they're short."
- I don't desire that if I were short then I be beat up.

I'd likely violate this and be inconsistent (especially if I *know* what it's like to be beat up and *imagine* myself, vividly and accurately, in the place of short people treated this way). This same consistency approach can help us to criticize other discriminatory principles (racial, religious, gender, sexual orientation, etc.).

Again we can derive this conscientiousness imperative using [p]:

1 All short people ought to be beat up, just because they're short. (Premise.)
∴ If I were short, then I ought to be beat up. (From premise 1.)
∴ If I were short, then beat me up. (From line 2 using [p].)
∴ Don't combine accepting "All short people ought to be beat up, just because they're short" and not accepting "If I were short, then beat me up." (From 1 to 3 using [r].)
∴ Don't combine believing "All short people ought to be beat up, just because they're short" and not desiring that if I were short then I be beat up. (This substitutes equivalents.)

Note the difference between these two questions:

- GOOD FORM: Do you *desire that if* you were short then you be beat up?
- BAD FORM: *If* you were short, *then would you desire* to be beat up?

We need "*desire that if*" (the first form) to show the person's *present inconsistency* (which involves his *present desire toward a hypothetical situation*). This point will come up again later, when we discuss how to formulate the golden rule.[23]

[23] During the Vietnam war, a pacifist friend of mine was asked by his draft board: "If killing a madman were the only way to stop him from killing your family, would you kill him?" I objected to my friend later that they should have asked: "Do you now desire that, if killing a madman were the only way to stop him from killing your family, then you wouldn't kill?" The latter question tests the pacifist's current consistency.

We discussed earlier (in section A) how to criticize Ima Racist's *arguments*, like "We ought to treat blacks poorly – *because* they're inferior." But suppose that Ima Racist appeals, not to an *argument*, but to a *basic principle* like this: "All people with dark skin ought to be treated poorly, just because they have dark skin." We could use a consistency challenge here too, just as in our shortism example:[24]

Don't combine these:

- I believe "All people with dark skin ought to be treated poorly, just because they have dark skin."
- I don't desire that if I had dark skin then I'd be treated poorly.

In rare cases, Ima may be *consistent* in holding his racist principle; he might *believe* that he ought to be treated poorly if he had dark skin – and *desire* to be treated that way in their place. Ima might do this because of:

- *Lack of knowledge*. Ima may believe that blacks don't suffer from being treated poorly; and so he can desire that if he were in their place then he'd be treated poorly too (since he wouldn't suffer from this). Then Ima needs more knowledge.[25]
- *Lack of imagination*. Ima may not have vividly and accurately visualized himself in the place of blacks being treated poorly; and so he can think that he desires to be treated poorly in their place. Then Ima needs more imagination.[26]
- *Perverted desire*. Ima may so hate the idea of being black that he hates himself when he imagines himself being black; so he desires

[24] Instead of appealing to consistency against Ima, we may be tempted to counter with our own principle, like "People of all races ought to be treated with respect." While this is a fine principle, Ima will just reject it. And so we'll have a stalemate, where Ima has his moral intuitions (or moral feelings) and we have ours, and neither can convince the other. A consistency appeal is more decisive, since it turns Ima's own principle against himself.

[25] Harriet Beecher Stowe's novel *Uncle Tom's Cabin* (New York: Harper & Row, 1965, first published 1852), p. 125, suggests that southern slaveowners often believed (but quite wrongly) that black slave parents weren't very disturbed when their children were sold off to make money for the slaveowners.

[26] Charles Darwin's *The Voyage of the Beagle* (New York: P.F. Collier, 1909, first published 1839) was amazed when his Brazilian host family sold off slave children to make money; he attributed this moral blindness to a lack of imagination: "Those who look with a cold heart at the slave never put themselves into the position of the latter. Picture your wife and little children being torn from you and sold like beasts to the first bidder!"

that if he were black then he be treated poorly.[27]

Perverted desires are the most interesting case. Such desires typically would come from a social conditioning that uses false or slanted beliefs. Maybe Ima Racist was taught that blacks are intellectually or morally inferior to whites. Maybe he was told only bad things about blacks and only good things about whites (even though both groups have both good and bad). Maybe his family and friends hated blacks, called them names, and promoted stereotypes about them. Maybe he met only a few atypically nasty blacks. Then Ima's hateful desires would diminish if he got his facts straight, understood the origin of his hatred, and broadened his experience of black people in an open way. So greater knowledge and experience would tend to diminish his hateful anti-black desires.

We also can have the racist consider other prejudices. All over the world, people in one group are taught to dislike those of another group. We teach young children: "Be suspicious of *those other people*. They're of a different race (religion, ethnic background, sexual orientation, or caste). They aren't our kind. They have strange customs and do strange things. They're evil and inferior." *People often believe very negative things about other groups on very flimsy evidence.* When we broaden our knowledge and personal experience, we conclude, "They're people too, much like us, with many of the same virtues and vices."

Hatreds programmed into us from our youth may never completely disappear; but a wider knowledge and experience will reduce them. That's all that our consistency arguments need. Only a very strong hatred of blacks can make us desire that if we were black (or found out to be black[28]) then we be treated badly. And we can criticize such desires on rational grounds.

---

[27] There's an allegedly true story about a Nazi who so hated Jews that he came to hate himself and his family when he discovered that they had Jewish ancestry. So he had himself and his family put into concentration camps and killed. This Nazi was consistent. But, I argue, his fanatical desires can be rationally criticized.

[28] The traditional American definition makes you *black* if you have *one* clearly black ancestor (even though you might be white by skin color, racial features, or culture). By this definition and DNA tests, about 10% of apparently white people in some former slave states are black. Such definitions make racist principles difficult to apply. The anti-racist idea that all people of all races and groups ought to be treated with respect and consideration (as we ourselves want to be treated) is easier to apply – since this applies regardless of how we arbitrarily divide the races.

E. Impartiality

Impartiality requires that we make similar evaluations about similar actions, regardless of the individuals involved. If we're impartial, we evaluate an action based on what the action is like – and not based on who plays what role in the situation. If we judge that an action is right (or wrong) for one person to do, then we judge that the same action would be right (or wrong) for anyone else to do in the same situation.

I violate impartiality if I make conflicting evaluations about actions I regard as exactly similar or relevantly similar. Two actions are *exactly similar* if they have all the same properties in common. They're *relevantly similar* if the reasons why one fits in a given evaluative category (good, bad, right, wrong, or whatever) also apply to the other. In the actual world, no two actions are ever exactly similar (have all the same properties in common). But the notion applies to hypothetical cases. To test my impartiality, I can imagine an exactly similar action where I'm on the receiving end of the action.

Here's my Good Samaritan example (Luke 10:30–35). Suppose that, while jogging, I see a man who's been beaten, robbed, and left to die. Should I help him, perhaps by running back to make a phone call? I think of excuses why I shouldn't; I'm busy, don't want to get involved, and so on. I say "It would be all right for me not to help him." But then I consider an exactly reversed situation. I imagine myself in his place; I'm the one beaten, robbed, and left to die. And I imagine him in my place; he's jogging, sees me, and has the same excuses. I ask, "Would it be all right for this man not to help me in this situation? Surely not!" But then I'm inconsistent. What's all right for me to do to another has to be all right for the other to do to me in an imagined exactly reversed situation.

In the actual world, no two actions are ever exactly similar. But we can always imagine an exactly similar action where I'm on the receiving end of the action. I violate impartiality if I violate this principle:

Don't combine these:

- I believe "It would be all right for me to do such and such to X."
- I believe "In an exactly similar situation, it would be wrong for X to do this to me."

This sounds a little like the golden rule. But it's about impartiality, about making similar *evaluations* about similar actions. The genuine golden rule is about *actions* and *desires* ("*Treat* others as you *want* to be treated"), not about evaluations.

My example uses an imagined "exactly reversed situation" where all my properties are switched with those of the other person. Let me explain this further. Suppose we list my properties and those of the other person (X):

My properties: jogging, very busy, has blue eyes, …

X's properties: beaten and robbed, needs a doctor, has brown eyes, …

Imagine that the list contains all our properties, even complex ones; the list would be too long to write out – perhaps infinitely long. When I imagine an exactly reversed situation, I imagine the list of properties being reversed:

My properties: beaten and robbed, needs a doctor, has brown eyes, …

X's properties: jogging, very busy, has blue eyes, …

Here I'm beaten and robbed, and X is the jogger. We also reverse relationships; so if X helped me in the past, I'd imagine that I helped X in the past.

Instead of switching all the properties in my mind, I can switch just the ones *relevant* to evaluating the action. If I'm not sure if a property is relevant, I can switch it anyway, to be safe. Then I imagine a "relevantly similar" situation.

Suppose I'm driving and see a hitchhiker. Should I pick him up? If I don't, he may spend a long time waiting; I know what this is like from when I've hitchhiked to backpacking trailheads. On the other hand, people who pick up hitchhikers are sometimes robbed or hurt. Impartiality tells me that whatever judgment I make on my picking up the hitchhiker (that it's obligatory, wrong, or neutral), I must make the same judgment on the imagined reversed-situation action. Impartiality doesn't tell me what to do; and here it doesn't push me toward an obvious answer. Rather, it encourages me to reflect on the action from both perspectives (mine and the hitchhiker's). And it insists that, whatever I decide, I must apply the same standards to myself that I apply to others.

These examples test our impartiality by seeing how we evaluate an *imagined* second case. The next example uses an *actual* second case that's recognized to be relevantly similar. This example also shows how such reasoning can help us to recognize duties toward ourselves.

In the movie *Babe*, the athlete Babe Didrikson (1914–56) needed a colostomy operation to save her life. From fear, she decided that she shouldn't have it. But her husband, thinking that she should have it, had her talk with another woman in similar circumstances, who also had to choose between dying and having the operation. Babe instinctively told the woman, "You ought to take courage and have the operation – for life is our greatest

gift." But then Babe realized that she had to apply the same principles to herself that she applied to others. So she decided that she too ought to have the operation.

Babe at first violated this consistency imperative:

Don't combine these:

(1) I believe "I ought not to have the operation."

(2) I believe "You ought to have the operation."

(3) I believe "Our cases are relevantly similar."

She held (1) because she feared the operation's results, she held (2) because she thought life is our greatest gift, and she held (3) because she thought that any reasons that justify one operation would also justify the other. Since her beliefs were inconsistent, she had to reject (1), (2), or (3). She in fact rejected (1), saying that she too ought to have the operation, just as the other woman in her similar situation should have it. She could have rejected (2), saying that neither should have the operation; she didn't do this, because she so strongly believed that the other woman ought to have it. Or she could have rejected (3), saying that the operation was right in one case but not the other, because of such and such differences; but she couldn't think of reasons that justify one operation but not the other. So consistency, while not telling her what exactly to believe, helped her to form her beliefs.

Some object that appealing to relevantly similar actions is slippery. What keeps you from appealing to trivial differences – from saying "It's all right for me to kill you but wrong for you to kill me, because I happen to have six toes and you don't"? If you pick trivial differences, we can appeal to hypothetical cases. Imagine a case where you have six toes (instead of me). Do you really think that then it would be all right for you to kill me? No one would believe this. In appealing to relevant differences, we have to give the factor equal weight regardless of which side is imagined to have it; this isn't easy to satisfy. At times, though, it's cleaner to appeal to imagined exactly similar cases.

Consistency norms respect our moral freedom, since they don't tell us specifically what to do or believe. They also promote moral rationality, since they guide us on how to work out our views consistently; they forbid, for example, criticizing others for doing certain things without also criticizing ourselves when we do the same things in similar circumstances.

I've spoken of impartiality as a type of "consistency" between our evaluations. I leave it open whether to take this as "logical consistency," or more generically as "uniformity." Suppose I make conflicting evaluations

about similar actions. Richard Hare's prescriptivism says that I violate logical consistency (since I misuse the term "ought"). Other views may say that I violate an impartiality duty to make similar evaluations about similar actions; this duty might perhaps rest on social conventions, personal ideals, divine commands, or self-evident truths. On these views, violating impartiality involves an objectionable clash between moral evaluations but not necessarily a logical inconsistency or self-contradiction.

Impartiality, as I use the term, requires that we make similar evaluations about actions we regard as exactly or relevantly similar; it doesn't say these things:

- "If it's all right for my sister to drive, then it's all right for me to drive." This needs a clause about the actions being relevantly or exactly similar; maybe your sister has a license but you're too young to have one.
- "Always act the same way in the same kind of situation." It's fine to eat carrots one day and celery another day in a similar situation, both actions being neutral (all right to do and all right to omit doing). We violate impartiality only when we make conflicting *evaluations* about actions we regard as relevantly or exactly similar.
- "Treat everyone the same." We may give more help to one who needs it (as in the Good Samaritan case) or prescribe different treatments for patients with different illnesses. These needn't involve making conflicting evaluations about similar actions.
- "Love everyone equally." This would destroy friendships and families. Suppose we love our children more that we love strangers – and we think it would be all right for any parent in similar cases to do the same thing. Then we're making similar evaluations about similar actions, and thus we satisfy impartiality.

On the last point, some forms of utilitarianism require *strong impartiality* ("We ought to have equal concern for everyone's good"). I require only *weak impartiality* ("We ought to make similar evaluations about similar cases"); this lets me believe that *I* ought to have greater concern for *my* children so long as I believe that in similar cases *others* ought to have greater concern for *their* children.

Our last axiom is [u] (for "universalizability"),[29] used to derive impartiality imperatives:

[u] From "Act A is all right (wrong, ought to be done)" derive "Any act B exactly or relevantly similar to act A in a second situation also is all right (wrong, ought to be done)."

Again, this is meant to leave open whether universalizability is a logical entailment. Here's an example of how to derive impartiality imperatives using [u]:

1 It would be all right for me to do such and such to X. (Premise.)

2 In an exactly similar situation, it would be wrong for X to do this to me. (Premise.)

∴ In an exactly similar situation, it would be all right for X to do this to me. (From premise 1 using [u], and it contradicts the previous line.)[30]

∴ Don't combine accepting "It would be all right for me to do such and such to X" and accepting "In an exactly similar situation, it would be wrong for X to do this to me." (This follows using [r] and the fact that 1 and 2 with [u] leads to a contradiction.)

∴ Don't combine believing "It would be all right for me to do such and such to X" and believing "In an exactly similar situation, it would be wrong for X to do this to me." (This substitutes equivalents.)

---

[29] I once read all the discussions of universalizability that I could find. I found 111 thinkers explicitly accepting universalizability, and no one explicitly rejecting it. (I didn't include those who reject formulas that I also reject.) So the consensus in favor of universalizability was 111 to 0. Philosophers debate, not so much the truth of universalizability, but rather its usefulness and justification. Here I try to show that it's very useful, but I'm neutral about how to justify it.

The precise technical wording for the "all right" part of [u] goes: "If act A is all right, then there's some universal property (or conjunction of such properties) F, such that: act A is F, and in any actual or hypothetical case any act that's F is all right." Here a universal property is a non-evaluative property describable without proper names (like "Gensler" or "Boston") or pointer terms (like "I" or "this"). An exactly reversed situation switches all the universal properties. See *Introduction to Logic* §§14.4–14.6 and *Formal Ethics* §§4.1–4.4.

[30] The precise technical analysis of "If it's all right for you to do A to X, then it would be all right for X to do A to you in the exact same situation" is this: "If it's all right for you to do A to X, then, for some universal property F, F is the complete description of your-doing-A-to-X in universal terms, and, in any actual or hypothetical case, if X's-doing-A-to-you is F, then it would be all right for X to do A to you." See *Introduction to Logic* §14.5.

When appealing to impartiality or universalizability, remember to include a *same-situation clause*.

F. The Golden Rule

The golden rule (GR) says "Treat others as you want to be treated." GR is a global standard, endorsed by nearly every religion and culture, important for families and professionals across the planet, and a key part of a global-ethics movement.[31]

Here's my favorite story to introduce GR.[32] There once was a grandpa who lived with his family. As Grandpa grew older, he began to slobber and spill his food; so the family had him eat alone. When he dropped his bowl and broke it, they scolded him and got him a cheap wooden bowl. Grandpa was so unhappy. Now one day the young grandson was working with wood. "What are you doing?" Mom and Dad asked. "I'm making a wooden bowl," he said, "for when you two get old and must eat alone." Mom and Dad then looked sad and realized how they were mistreating Grandpa. So they decided to let him eat with the family and to keep quiet when he spills his food.

The heart of the golden rule is switching places. You step into another's shoes. What you do to Grandpa, you imagine being done to you. You ask, "Am I willing that if I were in the same situation then I be treated that same way?"

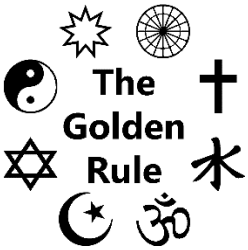
GR seems simple. But loose GR wordings invite objections; many academics dismiss GR as a folksy proverb that self-destructs when analyzed. I think we just need to understand GR better. I put my improved wording on a shirt.[33] It has "the golden rule" with symbols for eight GR religions (Bahá'í, Buddhism, Christianity, Confucianism, Hinduism, Islam, Judaism, and Taoism). It also has my GR formula, which tries to help us apply GR to difficult cases.

---

[31] To supplement my somewhat technical approach, I highly recommend Jeffrey Wattles's *The Golden Rule* (New York: Oxford, 1996), which emphasizes historical and religious aspects.

[32] Grimm, Brothers (1812), "The old man and his grandson," http://www.gutenberg.org/ebooks/2591. See also http://www.harryhiker.com/goldrule.htm (my GR page, or http://www.harrycola.com/goldrule.htm) and http://www.harryhiker.com/stories.htm (GR stories, or http://www.harrycola.com/stories.htm).

[33] See https://www.zazzle.com/store/harrygensler.



Treat others only as you consent to being treated in the same situation.

My GR formula commands consistency. It demands a harmony between my action toward another and my desire about how I'd be treated in the same situation. It doesn't replace other moral norms or theories, or give all the answers. It doesn't say specifically what to do (so it doesn't command bad actions if we have flawed desires). Instead of directly specifying right and wrong, it helps us figure this out for ourselves – by forbidding inconsistent combinations:

Don't combine these:

- I do A to another.
- I'm unwilling that if I were in the same situation then A be done to me.

GR, far from being vague, is a precise consistency test. Suppose I force Grandpa to eat alone. I switch places in my mind: I imagine that I'm forced to eat alone in the same situation. Do I condemn this same action done to me? Then I condemn how I treat Grandpa. I condemn how I treat another, if I condemn the same action when I imagine it done to me in the same situation.

To apply GR, imagine yourself in the other person's place on the receiving end of the action. You violate GR if you act in a given way toward another but are unwilling to be treated that way in the same situation. To lead reliably to right action, GR consistency needs to build on *knowledge* (of how your action affects another) and *imagination* (visualizing yourself in another's place).

We can derive GR using principles [p], [u], and [r]:

1 Do A to X. (Premise.)
| Assume: It would be wrong for me to do A to X. (Assumption for RAA argument).
| ∴ Don't do A to X. (From the assumption using [p]; it contradicts 1).

∴ It wouldn't be wrong for me to do A to X. (From lines 1 to 3 by RAA.)[34]
∴ It would be all right for me to do A to X. (From previous line.)
∴ In an exactly similar situation, it would be all right for A to be done to me. (From previous line using [u].)
∴ In an exactly similar situation, A may be done to me. (From previous line using [p].)
∴ Don't combine accepting "Do A to X" and not accepting "In an exactly similar situation, A may be done to me." (This follows using [r] and the fact that 1 with [p] and [u] entail the previous line.)
∴ Don't combine acting to do A to another and being unwilling that if I were in the same situation then A be done to me." (This substitutes equivalents.)

So GR on my approach isn't basic; instead, it's a theorem provable using [p], [u], and [r].[35]

G. Three Technical GR Distinctions

Our GR formula has three refinements that we needed before:

- a *similar situation* qualifier (section E),
- a *present attitude to a hypothetical situation* (section D, *desire-that-if*), and
- a *don't-combine* form (section A).

Without these, GR has absurd consequences and can't be proven from our axioms.

Three parts of my GR formula are crucial. (1) The first part is "*in the same situation*."

Don't combine these:

- I do A to another.
- I'm unwilling that if I were *in the same situation* then A be done to

[34] Acting (accepting "Do A") commits us to believing that the act is permissible (accepting "Act A is all right"). While this could be built into [p], I prefer to keep [p] simpler and derive this area using an RAA (reductio ad absurdum) argument.

[35] *Introduction to Logic* ch. 14 symbolizes GR as "~(u:Aux • ~u:(∃F)(F*Aux • □(FAxu ⊃ MAxu)))" – which can be paraphrased "Don't combine (1) accepting 'Do A to X' with (2) not accepting 'For some universal property F, F is the complete description in universal terms of my-doing-A-to-X, and, in any actual or hypothetical situation, if X's-doing-A-to-me is F, then X may do A to me.'" The proof takes 35 steps.

me.

People who reject GR usually understand it crudely, often as what I call the *literal GR* ("*If you want X to do A to you, then do A to X*"). By this, if you want Lucy to be kind to you, then be kind to her; and if you want Adam not to hurt you (or rob you, or be rude to you), then don't do these things to him. These seem sensible. The literal GR generally works well if you and X are in similar situations and if you have good desires about how you're to be treated. But if either condition fails, the literal GR can tell you to do crazy or evil things:

- *Different situations*. "If you want your doctor to remove your appendix, then remove your doctor's appendix."
- *Flawed desires*. "If you want others to hurt you [suppose you do], then hurt them."

We can avoid these problems by wording GR more carefully.

First, you and the other person may be in very different situations. Consider this instance of the literal GR:

To a son who hears
well but has a father
with bad hearing: If
you want your father
not to speak loudly to
you [since your hearing
is normal], then don't
speak loudly to him.

This ignores differences between you and your father; a same-situation clause fixes the problem. Ask this: "How do I desire that I'd be treated if I were in the same situation as my father (and thus hard of hearing)?" You desire that if you were in his same situation then people would speak loudly to you; so you speak loudly to him.

We can take "same" situation here as "exactly similar" or "relevantly similar." In the first case, I imagine myself in my father's exact place (with all his properties). In the second, I imagine myself having those properties of my father (such as being hard of hearing) that I think are or may be relevant in deciding how to speak to him; if you're unsure whether a property is relevant, then switch it anyway, just to be safe. Both approaches work fine.

Here's another case where the literal GR leads to problems:

To a patient: If you
want your doctor to
remove your
appendix, then
remove your
doctor's appendix.

Again, we need a same-situation qualifier. The patient clearly doesn't desire that if he were in his doctor's place (with a healthy appendix), then his appendix be removed by a sick patient ignorant of medicine. As you apply GR, ask this: "Am I willing that if I were in the same situation then this be done to me?" The other person's situation includes likes and dislikes. So if you're a waiter who hates broccoli, but your customer likes and orders it, then you imagine being served broccoli in a hypothetical situation where you like and order it.

We need a same-situation qualifier because people may have different needs and circumstances (and because GR is based on impartiality, which needs this same qualifier). The literal GR ignores such differences and has us treat everyone the same way, in accord with *our* needs and circumstances.[36]

The *literal GR fallacy* assumes that everyone has the same likes, dislikes, and needs that we have. Here's a story to illustrate the fallacy. There once lived a monkey and a fish. The monkey followed GR, always trying to treat others as he wanted to be treated. But he sometimes applied GR foolishly. Now one day a big flood came. As the threatening waters rose, the monkey climbed a tree to safety. He looked down and saw a fish struggling in the water. He thought, "I wanted to be lifted from the water." And so he reached down and grabbed the fish, lifting him from the water to safety on a high branch. Of course that didn't work. The fish died.

The monkey applied GR literally: treat others as you want to be treated. He wanted to be lifted from the water, so he lifted the fish from the water. He

[36] The literal GR can prescribe contradictory actions. Suppose that you want Alice and Betty to do whatever *you* ask; so then you are to do whatever *they* ask. Alice asks you to do A, and so you are to do A. Betty asks you *not* to do A, and so you are *not* to do A. So facts plus the literal GR can entail contradictions.

Can we avoid the literal-GR absurdities by applying GR only to "general" actions (like treating someone with kindness) – and not "specific" actions (like removing someone's appendix)? This won't work, since hurting someone or doing what others ask are surely *general* actions, and yet they lead to problems.

didn't consider how monkeys and fish differ. Being lifted from the water saves a monkey but kills a fish. So the monkey applied GR foolishly.

There was also a wise GR monkey, named Kita, who considered lifting a fish out of the water. But Kita knew that this would kill the fish. As she imagined herself in his situation, she asked, "Am I now willing that if I were in the same situation as the fish, then I be lifted from the water?" She answered, "Gosh no: this would kill me!" So she left the fish in the water. Since people (and animals) may differ, this same-situation clause is crucial; we are to treat others only as we consent to being treated *in the same situation*.[37]

"KITA" is also an acronym (Know-Imagine-Test-Act) for some elements for using GR wisely:

- Know: "How would my action affect others?"
- Imagine: "What would it be like to have this done to me in the same situation?"
- Test for consistency: "Am I willing that if I were in the same situation then this be done to me?"
- Act toward others only as you're willing to be treated in the same situation.

To lead reliably to right action, GR consistency needs to build on things like knowledge, imagination, creativity, rationalized desires, and a healthy self-love.[38]

---

[37] *Ethics and the Golden Rule*, ch. 5 has a brief GR history in the form of a date-event chronology. (See also http://www.harryhiker.com/chronology.htm or http://www.harrycola.com/chronology.htm.) The earliest written GR sayings that we have go back to about the middle of the first millennium BC, in China, India, Persia, and Greece, with the emergence of written language. The first objection to the literal GR that I could find goes back to Augustine (354–430) and the first use of a same-situation clause that I could find goes back to Francis of Assisi (c. 1220). Later, a number of English GR discussions used same-situation clauses, including discussions by Benjamin Camfield (1671), George Boraston (1684), John Goodman (1688), and Samuel Clarke (1706). In explaining how to express GR so that it doesn't lead to absurdities, it's wise to start with the need for a same-situation clause.

[38] GR works best as part of a team; it carries the ball, but it needs other team members to block. GR appeals to factors like understanding, imagination, and desires; any of these may be flawed and criticized further. This is possible because GR – as a *don't-combine imperative* instead of an *if-then* – doesn't by itself directly tell us what to do; instead, it gives a consistency condition that we need to fulfill (but we need to fulfill other conditions too, such as being informed and imaginative).

My favorite historical GR example is a civil-rights speech by President John Kennedy (11 June 1963),[39] during the first black enrollment at the University of Alabama. While Kennedy didn't know about GR monkeys, his speech followed KITA:

- He first got people to *know* how blacks were being treated, in areas like education, employment, and voting.
- He had whites *imagine* themselves being treated that way on the basis of their skin color.
- To *test* their consistency, he asked whether they'd be content to being treated that way.
- He urged *acting* on GR: "The heart of the question is whether all Americans are to be afforded equal rights and equal opportunities, whether we are going to treat our fellow Americans as we want to be treated." He called for changes in actions, attitudes, and laws (including the 1964 Civil Rights Act).

The heart of morality is GR. And the heart of GR is switching places. What we do to Grandpa (or blacks, gays, or whomever we mistreat) we imagine being done to ourselves. And to avoid the literal GR fallacy, we can imagine ourselves in the other's exact place (having their likes, dislikes, needs, and so on).

(2) The "*unwilling that if*" part of my GR formula is also very important:

Don't combine these:

- I do A to another.
- I'm *unwilling that if* I were in the same situation then A be done to me.

GR is about our present reaction to a hypothetical situation; it isn't about how we'd react if we were in that situation.

Suppose you're a nurse about to give a shot to a baby. The baby doesn't want the shot but needs it to neutralize a bee sting that would otherwise bring great pain or death. To apply GR, you imagine yourself in the baby's place. You imagine this situation: You're a baby who doesn't want the shot but needs it to neutralize a bee sting that would otherwise cause great pain or death. What do you now want to happen in this situation? Clearly, you'd now say "Give me the shot." So you're now willing that if you were in this situation then you'd be given the shot (even against your will). GR has you ask this question: "Am I now *willing that if* I were in the same situation as

[39] Https://www.americanrhetoric.com/speeches/jfkcivilrights.htm.

this baby then I be given a shot?" The answer is *yes*. So GR, correctly understood, lets you give the baby the shot.

People often ask the GR question wrongly, which forces them to do whatever the other person wants. They ask, "If I were in the baby's place, how would I then want to be treated?"[40] Now if you were in the baby's place, then you wouldn't want to be given the shot; misapplying GR, we'd conclude that you shouldn't give the baby the shot. So it's better to apply GR as explained above. I can give the baby the shot (to protect it from great pain or death), since I'm now willing that if I were in her situation then I be given the shot. In asking the GR question, it's important to say "willing that if":

Am I *willing that if*
I were in the same
situation then this
be done to me?

Immanuel Kant's objection to GR rests on this confusion. Here you're a judge, about to sentence a dangerous criminal to jail. The criminal protests and appeals (incorrectly) to GR: "If you were in my place, you wouldn't want to be sent to jail; so by the golden rule you can't send me to jail." You should respond: "I can send you to jail, because I'm now *willing that if* I were in your place (as a dangerous criminal) then I be sent to jail." You could add, "If I do such things, then please send me to jail too!"[41]

Sometimes we need to act against what others want. We may need to give a shot to babies who don't want it, refuse salespersons who want to sell us overpriced products, fail students who don't work, defend ourselves from attackers, or jail dangerous criminals. GR lets us act against what others want, if we're now willing that if we were in their situation then we be treated similarly.[42]

---

[40] Asking the question the wrong way turns GR into the *platinum rule*: "Treat others as THEY want to be treated." Since the little baby wants not to be given the shot, you wouldn't give her the shot (and so she'd get great pain or die).

[41] *Groundwork of the Metaphysics of Morals*, trans. H.J. Paton (New York: Harper & Row, 1964), p. 97; for more on how I answer Kant's GR criticisms, see my *Ethics: A Contemporary Introduction* §7.11 and *Ethics and the Golden Rule* §14.3c. My *Ethics and the Golden Rule* answers 33 objections to GR. For example, 30 objects that GR appeals to selfish self-interest while 31 objects that GR appeals to an unrealistic pure altruism; both are wrong – GR says nothing about motivation and is compatible with either self-interest or altruistic motivation.

[42] All these cases provide objections to the *platinum rule* ("Treat others as THEY want to be treated"). GR would push us to take some account of the desires of others, in most cases, but less in the cases mentioned here.

(3) The "*don't combine these*" part of my GR formula is also very important.

Don't combine these:

- I do A to another.
- I'm unwilling that if I were in the same situation then A be done to me.

GR forbids an inconsistent combination; it doesn't say which individual actions are right or wrong.

Recall that the literal GR can lead to absurdities in two main ways. We dealt with the first (different-circumstances) problem by adding a same-situation clause. A second problem is that the literal GR can tell us to do bad things if we have flawed desires about how we're to be treated. I'll give four examples.

(A) There once was a woman named Electra. Electra wanted to follow GR, but she got her facts wrong; she thought electrical shocks were pleasant. Since she wanted others to shock her, she applied the literal GR and shocked them: "If you want others to give you electrical shocks, then give them electrical shocks." Given flawed desires, the literal GR can command evil actions.

We'll use a triple defense against flawed desires. (i) My improved GR formula, instead of telling us specifically what to do, just forbids a combination:

Don't combine these:

- I give electrical shocks to another.
- I'm unwilling that if I were in the same situation then electrical shocks be given to me.

Since my improved GR doesn't say specifically what to do, it doesn't tell Electra to do evil things (like shock others).

(ii) GR consistency, to lead reliably to right action, needs to build on other things, like knowledge and imagination. If we're misinformed, then we might do evil things while satisfying GR consistency. Here Electra shocks others (an evil thing) but satisfies GR consistency (she's willing that she be shocked in similar cases), since she's misinformed and thinks these shocks are pleasurable.

(iii) We need to use reason against flawed desires. Here we'd show Electra that electrical shocks are painful (perhaps by giving her a small one).

Once she understands this, GR consistency will lead her away from shocking others.[43]

(B) Or suppose Mona hates herself and wants others to hate her; then the literal GR tells her to hate others. (i) But again, the correctly formulated GR forbids a combination but doesn't tell her to hate others. (ii) GR consistency, to lead reliably to right action, needs to build on other things (like knowledge, imagination, and here a healthy self-love). (iii) We can use reason against Mona's flawed desires. We can try to help Mona understand why she hates herself and how to neutralize this hatred – by not fixating on her negatives, by seeing herself and her good points in a more balanced way, and, if she's a believer, by appreciating how God loves her. Once Mona regains a healthy self-love, GR consistency will lead her more readily to love others.

(C) Or suppose Mike is a masochist who gets athletic satisfaction from pain and wants others to cause him pain; then the literal GR tells him to cause pain to others. My improved GR would handle this much like how it handles the Mona case. But here the same-situation clause is also important, since Mike is likely unwilling that he be caused pain if he were in the place of his victims (who presumably get no satisfaction from pain).

---

[43] Even with a same-situation clause and a present attitude toward a hypothetical situation (desire-that-if), GR formulated as an *if-then imperative* can command evil or contradictory actions. Consider this *if-then imperative* formula:

- Electra, if you desire that if you were in X's exact place then you be given a severe electrical shock, then give X a severe electrical shock.

Imagine that Electra, as in the text, has defective beliefs and desires about electric shocks, and thus satisfies the if-part. Then this *if-then imperative* GR tells her to do an evil action. It's very easy to multiply further examples.

In 3-party GR applications, a similar *if-then imperative* GR can easily command contradictory actions. Imagine that your friend Alice wants you to help her to rob Betty – and you get conflicting results depending on whether you imagine yourself in Alice's place or Betty's place:

- If you desire that if you were in Alice's exact place then you be helped to rob Betty, then help Alice to rob Betty.
- If you desire that if you were in Betty's exact place then Alice not be helped to rob you, then don't help Alice to rob Betty.

Suppose that both if-clauses are true; then the *if-then imperative* GR would tell you to do contradictory things: "Help Alice to rob Betty" and "Don't help Alice to rob Betty." We avoid such problems by moving to *don't-combine* formulations; see section H for how to deal with such cases.

(D) Or suppose Adolf is a Nazi who so hates Jews that he desires that he be killed if he were found to be Jewish; then the literal GR tells him to kill others if they're found to be Jewish. Again, we can make three points. (i) The correctly formulated GR forbids a combination but doesn't tell Adolph to kill Jews. (ii) GR consistency, to lead reliably to right action, needs to build on other things (like knowledge, imagination, and here rational desires). (iii) We can use reason against Adolph's flawed desires. When we try to understand why he hates Jews so much, we'll likely find that his hatred has its source in things that can be rationally criticized. Maybe Adolf thinks Aryans are superior to Jews and racially pure; we can criticize this on factual grounds. Or maybe Adolf was taught to hate Jews by his family and friends, who hated Jews, called them names, and spread false stereotypes about them. Then his anti-Jewish desires came from false beliefs and social conditioning; his flawed desires would diminish if he understood their origin and broadened his experience and knowledge of Jews in an open and personal way. With greater rationality, Adolf wouldn't desire that he'd be killed if found out to be Jewish – and GR would be a powerful tool against his racism.

While this example was about a Nazi, the same idea applies to those who desire that they be mistreated if they were black, female, gay, or whatever. Such desires are likely flawed (as based on a social conditioning that uses false beliefs and stereotypes) and would be given up if we expanded our knowledge and experience of the group in an open and personal way.

The related *easy GR fallacy* assumes that GR gives an infallible test of right and wrong that takes only seconds to apply. Imagine a rich coal-mine owner named Rich, who pays his workers only $1 a day. He's asked if he's willing to be paid only $1 a day in their place. He replies, "Yes, you can live well on $1 a day, so I'm willing that I be paid that in the place of my workers; I love GR – you just whip out this moral compass and in a few seconds you know whether you're acting rightly!"

Rich moves too fast. To lead reliably to right action, GR needs to build on knowledge and imagination, which may take time. Rich is willing that he be paid $1 a day in his workers' place (and so is consistent), but he's so willing because he thinks (wrongly) that his workers can live well on this much. If he knew how little $1 buys, he wouldn't think this. Rich needs to get his facts right. He could begin by trying to go to the store to buy food for his family with only $1 in his pocket!

Now suppose that Rich decides to apply GR more adequately to how to run his mine. What would he do? Following KITA, he'd do four things.

- Rich would gain *knowledge*. He'd ask, "How are my company policies affecting others – workers, neighbors, customers, and so on?" To know this, Rich would need to spend time talking with workers and others.
- Rich would apply *imagination*. He'd ask, "What would it be like to be in the place of those affected by these policies?" He'd imagine himself as a worker (laboring under bad conditions for a poor salary), or a neighbor (with black smoke coming into his house). Or he'd imagine his children being brought up under the same conditions as the workers' children.
- Rich would *test* his consistency by asking: "Am I now willing that if I were in the same situation (as my workers, neighbors, or customers) then I be treated that same way?" If the answer is no, then his actions clash with his desires about how he'd be treated in a similar situation – and he must change something. To change company policies, Rich may need creativity to find alternatives. He might listen to ideas from others. He might learn what other companies and cultures do. He might imagine what policies make sense from a worker's perspective, explain current policies to a child, write an essay listing options, take a long walk, or pray about it. While GR doesn't tell him which policies to consider, GR gives a fairness test for any proposed policy; Rich must be able to approve of it regardless of where he imagines himself in the situation: as owner, worker, neighbor, or customer. The final solution, while maybe not satisfying everyone fully, needs to be at least minimally acceptable from everyone's situation.
- Rich would *act* on GR: "Treat others only as you consent to being treated in the same situation." Yes, it's a simple formula. But applying it wisely may require preparatory work on knowledge, imagination, and creativity. We'll never be perfect in these areas; but the fact that we may never do something perfectly doesn't excuse us from trying to doing it as well as we reasonably can.

The related *too-simple-or-too-complex GR fallacy* assumes that GR is either so simple that our kindergarten GR is enough for adult decisions or so complex that only a philosopher can understand it. On the contrary, GR is *scalable*. You can teach GR to small children ("Don't hit your little sister – you don't want us to hit you, do you?"), while adults can use it in complex decisions (like how to run a coal mine or a country in a way that respects

everyone's rights and interests). Our understanding of GR needs to grow as we mature.

People sometimes describe a situation to me and ask, "What would GR tell us to do in this situation?" This shows a misunderstanding of how GR works. Properly understood, GR doesn't directly tell us what to do, and so it doesn't give us the solution to our problem. We've got to propose our own solution (after studying the facts and imagining ourselves in the place of the various parties). Then GR can test our proposed solution for consistency (and fairness); any proposed solution must be one that we're willing to have followed regardless of where we imagine ourselves in the situation.

As a consistency norm (forbidding a combination but not directly telling us what individual action to do), GR isn't a direct criterion of right and wrong; and so GR isn't a rival to moral norms like "One ought not to steal." GR works at a different level. GR doesn't impose a rule on us from the outside but rather takes our own rule (e.g., "Don't steal from me!") and pushes us to apply it consistently to how to treat others. GR is much like "Don't contradict yourself." GR's role isn't to replace other ethical theories but to supplement them – by giving a consistency tool that's often useful. Most ethical theories recognize the role of consistency and so should be able to accept GR on this basis.

The golden rule, with roots in a wide range of world cultures, is well suited to be a standard that different cultures could appeal to in resolving conflicts. As the world becomes more and more a single interacting global community, the need for such a common standard is becoming more urgent.

GR is golden (valuable) because it captures so much of the spirit behind morality. It counters self-centeredness and helps us to see the point of moral rules. It's psychologically sound and personally motivating, engaging our own reasoning instead of imposing answers from the outside. It promotes cooperation and mutual understanding. It criticizes culturally taught racist or sexist moral intuitions or moral feelings. It concretely applies ideals like fairness and concern. And it's a global wisdom, common to most religions and cultures. So GR makes a good one-sentence summary of morality.[44]

---

[44] GR applies nicely to practical areas like racism, sexism and other forms of discrimination, global warming, moral education, how to treat animals, immigration, medical ethics, and business ethics; see *Ethics: A Contemporary Introduction*, pp. 124– 9, and *Ethics and the Golden Rule*, pp. 108–62. *Ethics and the Golden Rule* and *Formal Ethics* discuss many more issues about GR, like "Can we apply GR to how we treat animals?" "Are the positive and negative GRs ('Treat others as you want to be treated' and 'Don't treat others as you want not to be treated')

H. GR Variations

Even though I often speak of "*the* golden rule," this phrase is misleading – since GR is a *family* of related formulas instead of a single formula. So far, I've focused on "Treat others only as you consent to being treated in the same situation" – which is useful for moral thinking and can be derived from our four axioms; but many other wordings are also useful and can be derived in the same way. So our GR formula might imagine our daughter (or someone else we very much care about) on the action's receiving end. Or it might end "in an *exactly similar* imagined situation" or "in any *relevantly similar* actual situation." It might specify a duty: "You *ought* to treat others only as you consent to being treated in the same situation." It might deal with desires: "Don't combine *desiring* something to be done to another with being unwilling that this be done to you in the same situation." It might give a consistency condition for using moral terms: "Don't combine believing that it would be *all right* for you to do A to X and being unwilling that A be done to you in the same situation" or "Don't combine believing that you *ought* to do A to X with not wanting A to be done to you in the same situation." Since these and other variations can combine, there are at least 6,460 correct GR formulas.[45] So "the" golden rule is a family of principles, not a single principle. And there are related principles, like *self-regard*, *future-regard*, and the *generalized GR*.

Self-regard says: "Treat yourself only as you're willing to have others (especially those you most care about) treat themselves in the same situation." This can help us to recognize duties to ourselves. Maybe we have so much concern for our children that we never think of our own needs; we're inconsistent if we aren't willing that our children live that way when they grow up. Or we develop lazy work habits in college; we're inconsistent if we don't consent to the idea of a son of ours doing this. Or, because we lack courage and a sense of self-worth, we refuse to seek treatment for a drug habit that's ruining our life; we're inconsistent if we aren't willing that our younger sister do this in a similar situation.

---

significantly different?" "How does GR relate to 'Love your neighbor as yourself' and similar norms (like 'Treat others as brothers and sisters' or the Rawlsian 'Treat others only in ways that you'd support if you were informed and clear-headed but didn't know your place in the situation')?"

[45] See my *Formal Ethics*, pp. 101–4. There are many additional bad forms, which may lack a similar situation qualifier, a present attitude to a hypothetical situation, or a don't-combine form.

Many people have too little concern for themselves. Various factors (laziness, fear, habit, lack of self-appreciation, lack of discipline, and so forth) can drive us into actions that benefit neither ourselves nor others; consider how we hurt ourselves by overeating, selfishness, laziness, or overwork. Being *selfless*, if this means *having no concern for oneself*, is a vice. Our consistency norms recognize the importance of concern both for others and for ourselves.

Future-regard says: "Treat your future-self only as you're willing to have been treated by your past-self in the same situation." Here we switch times (imagining that we now experience future consequences) instead of switching persons. More crudely: "Don't do what you'll later regret." Maybe our drinking will cause a future hangover; but, when we imagine ourselves experiencing the hangover now, we don't consent to the idea of our having treated ourselves this way. Or our robbery will cause our future jail sentence; but when we picture ourselves being in jail now because of our past actions, we don't consent to our acting in a way that has such results. In both cases, we're inconsistent and violate future-regard.[46]

The generalized GR says: "Act only as you're willing for anyone to act in the same situation, regardless of where or when you imagine yourself or others."[47] This includes GR (where you, or someone else that you very much care about, is in the place of someone affected by your action), self-regard (where someone you care about is in your place), and future-regard (where you're at a future time experiencing your action's consequences). It also includes a multiparty GR, which has us satisfy GR toward all affected parties. Here's the *don't-combine* form and its derivation:

[46] Practical reason has a more unified structure than many philosophers realize. J.L. Mackie (*Ethics: Inventing Right and Wrong* [London: Penguin, 1977], pp. 228–9) contrasts three seemingly unrelated kinds of practical rationality: ends-means rationality, concern for yourself (especially your future), and concern for others; he complains that there is little unity among the three. Formal ethics instead finds a great unity.

[47] In a Kantian spirit, we might call the generalized GR the *principle of autonomy*. This means, not that you are to do as you feel like, but that you are to regulate your life by principles that you accept as holding for everyone equally and that you will to be followed regardless of where you imagine yourself in the situation.

Don't combine these:

- I do A.
- I'm not willing that every similar action be done (regardless of where or when I imagine myself or others in the situation).[48]

1 Do A. (Premise.)
| Assume: It would be wrong for me to do A. (Assumption for RAA argument).
| ∴ Don't do A. (From the assumption using [p]; it contradicts 1).
∴ It wouldn't be wrong for me to do A. (From lines 1 to 3 by RAA.)
∴ It would be all right for me to do A. (From previous line.)
∴ In every exactly similar situation, it would be all right for A to be done. (From previous line using [u].)
∴ In every exactly similar situation, A may be done. (From previous line using [p].)
∴ Don't combine accepting "Do A" and not accepting "In every exactly similar situation, A may be done." (This follows using [r] and the fact that 1 with [p] and [u] entail the previous line.)
∴ Don't combine acting to do A and not being willing that A be done in every exactly similar situation." (This substitutes equivalents.)

The generalized GR emphasizes that it's important to keep third parties in mind as we consider what to do. Suppose I own a store and need to hire just one worker. Alice and Betty apply, and I must choose between them. Here I must satisfy GR toward each party. So if I pick Alice (who's more qualified) instead of Betty, then I must be willing that I not be picked if I were in Betty's situation.[49] Combining two GRs, we get a three-party GR:

Don't combine these:

- I do A to X and Y.
- Either I'm not willing that A be done if I were in the place of X, or I'm not willing that A be done if I were in the place of Y.

[48] *Introduction to Logic* §14.5 symbolizes this as "~(u:Au • ~u:(∃F)(F*Au • □(X)(FX ⊃ MX)))" – which can be paraphrased "Don't combine (1) accepting 'Do A' with (2) not accepting 'For some universal property F, F is the complete description in universal terms of my doing A, and, in any actual or hypothetical situation, any act that is F may be done.'"

[49] The literal GR can lead to a contradiction when applied to the Alice-Betty case. By the literal GR, you should hire Alice (supposing that if you were in Alice's place you'd want to be hired) and you shouldn't hire Alice (supposing that if you were in Betty's place you'd want to be hired instead of Alice).

The generalized GR extends this to any number of affected parties: we must be willing that the act be done regardless of where we imagine ourselves in the situation. The affected parties may include future generations; this leads to the carbon rule: "Keep the earth livable for future generations, as we want past generations to have done for us."[50]

Some philosophical GR reconstructions don't sound much like the golden rule – like Alton's "If A is rational about rule R, then if there are reasons for A to think R applies to others' conduct toward A, and A is similar to those others in relevant respects, then there are reasons for A to think R applies to A's conduct toward others."[51] My philosophical GR reconstruction sounds much more like the ordinary golden rule: "Treat others only as you consent to being treated in the same situation." Like the usual GR formulas, this is an imperative and involves your action toward another and your desire about how you are to be treated. When I wear this on a t-shirt, as I often do (see section F), people instantly see it as a rewording of the golden rule. It adds a same-situation clause; but most people see this as a friendly clarification and nod in approval.[52] I've taught ethics at Loyola of Chicago to many very diverse groups (Christians, Muslims, Jews, Hindus, Buddhists, Bahais, Sikhs, non-believers, etc.) – and in China at the University of Wuhan[53] – and all of my students tend to see my formula as a cleaned up version (logically clearer and more defensible) of an idea deeply rooted in their own tradition. If you don't see the need to make the usual GR wordings logically clearer and more defensible, do a Web search for "problems with the golden rule"; if you

[50] *Ethics: A Contemporary Introduction* §8.5 has a longer application of GR to global warming, emphasizing how to answer climate-change deniers who say that, since there's no proof that human activity is the major cause for current temperature increases (which may take place for other random causes), thus we needn't change our use of fossil fuels.

[51] See Bruce Alton's *An Examination of the Golden Rule*, 1966 PhD dissertation at Stanford, http://disexpress.umi.com. This is largely historical, well worth reading, and is to my knowledge the first philosophy dissertation ever on the golden rule. The second such dissertation was my *The Golden Rule*, 1977 PhD dissertation at Michigan, http://disexpress.umi.com; this gives an earlier sketch of the ideas presented in this paper.

[52] This t-shirt version of my GR isn't as clear on the *willing-that-if* and *don't-combine* features, so in serious discussions it's wise to move quickly to my second formulation, which makes these clearer but is still pretty intuitive:

Don't combine these:

- I do A to another.
- I'm unwilling that if I were in the same situation then A be done to me.

[53] See http://www.harryhiker.com/china or http://www.harrycola.com/china.

understand this present paper well, you should be able to answer all of the objections raised by the many Web pages attacking the golden rule.

I. Metaethics

Formal ethics clarifies and systematizes a group of widely accepted, commonsense ethical sayings. My system rests on four axioms: [r], [e], [p], and [u]; taken together, these tell us to think and live consistently with logic and three further principles (about ends and means, keeping our moral beliefs in harmony with our lives, and making similar evaluations about similar cases). The theorems that follow are analogues of popular ethical sayings (like the golden rule and "Practice what you preach"), but they're formulated carefully to avoid absurd implications. Formal ethics is about being consistent; wisely applying its principles (like the golden rule) requires, in addition, knowledge and imagination. The whole system should be acceptable to a wide range of metaethical views about the nature of value. Here I'll give a few examples of such views and how they could accept my approach to formal ethics.

Intuitionism. The axioms and theorems of formal ethics are *self-evident truths* (ethical truths that strike most of us as clearly true and continue to strike us this way as we investigate them further). Many thinkers argue that moral ideas can't be self-evident, because they're vague and widely disputed; but these formal axioms and theorems are clear and have a strong consensus behind them. The norm to avoid inconsistencies is presupposed in practically every area of thought, including science and math. GR is widely held too, across different religions and cultures, and can be expressed in a clear way that resists objections. The appeal to consistency (including GR consistency) gives ways to dispute racist moral intuitions that our society may have taught us. And so an intuitionism based on formal ethics is the best form of intuitionism. (This, by the way, is the view that I accept.)

Cultural Relativism. Norms are social conventions; "good" means "socially approved" (by a given culture). While ethical systems vary a lot, there's some overlap. Except for perhaps very primitive societies, practically every society accepts the general ideas behind formal ethics; among these, the golden rule is especially important and widely accepted.[54] These general ideas are widely shared because they lead to social harmony, progress,

[54] The wide global presence of GR and related values has a biological and evolutionary explanation; see my *Ethics: A Contemporary Introduction*, pp. 161–70, and *Ethics and the Golden Rule*, pp. 125–35.

greater moral rationality, and a better life for all – which is why societies construct value systems. So the axioms and theorems of formal ethics are to be accepted as giving more precise and defensible formulations of widely shared social conventions.[55]

Emotivism and subjectivism. Morality is about your feelings (not about objective truths); "X is good" means "Hurrah for X!" (emotivism) or "I like X" (subjectivism). The norms of formal ethics can be based on your feelings in either of two ways. (1) If your feelings are *idealistic*, then you likely have positive feelings about being consistent (including GR-consistent); then you can accept these norms because they accord with your feelings. (2) If instead you're *egoistic*, then you can justify these norms by appealing to self-interest:

- Inconsistencies leads to confusion and frustrated desires. To see this, imagine how miserable you'd be if whenever you believed or wanted something you also believed or wanted the opposite. You'd go crazy! Likewise, false beliefs and lack of imagination also lead to confusion and frustrated desires.
- Inconsistency brings a stressful condition that psychologists call *cognitive dissonance*. Evolution programmed our minds to avoid inconsistencies.
- Inconsistencies and false beliefs bring social penalties. They cut us off from rational discussion and lead people to dismiss our ideas. Society is especially harsh when we violate conscientiousness or impartiality; it trains us to feel guilt, anxiety, and the loss of self-respect over this.
- People mostly treat us as we treat them. So it generally pays in terms of self-interest to treat others well, as we want to be treated.
- GR promotes cooperation, which benefits everyone (including us) and brings social rewards. Selfishness promotes conflict, which hurts everyone (including us) and brings social penalties. It's in everyone's self-interest to help bring about a social environment in which it's in no one's self-interest to violate GR, and this social environment somewhat exists around us.
- GR ennobles us. It enhances our sense of *self-worth*, which is essential to happiness. We see ourselves as having worth because that's how we see everyone. If we violate GR, then we lose self-respect; if we treat others as worthless, how can we see ourselves,

[55] See my "Values and Cultures," which will soon be published in Sanjit Chakraborty's *Minds and Cultures* (New York: Routledge, 202?).

who are cut from the same cloth, as having worth?

- Students volunteers typically report that they receive from others much more than they give to others. This gives evidence that helping others helps to make *us* happy.
- A simple experiment by Rimland[56] provides evidence that altruistic living promotes our happiness. Groups were asked to list persons they knew well and label them as happy-or-unhappy, and as altruistic-or-selfish. When responses are analyzed, "happy" people are almost always "altruistic" and almost never "selfish." So, judging by people's perceptions, self-interest supports GR.

So our feelings (whether idealistic or egoistic) can justify the norms of formal ethics.

Richard Hare's prescriptivism. This works like emotivism except that it sees conscientiousness, impartiality, and GR-consistency as built into our moral terms (and thus a matter of logical consistency) – and it sees norms of formal ethics as expressing our prescriptions (or desires) instead of our feelings.

Ideal-observer theory. "X is good" is a claim about what we'd desire under ideal conditions. It makes sense to include *being consistent* (as analyzed by formal ethics) as one of these ideal conditions – along with being informed and imaginative. Then formal ethics would provide part of the definition of "good."

Divine-command theory. "X is good" means "God desires X." The norms of formal ethics can be justified as mirroring God's desires. The Christian Bible often condemns inconsistency; so Jesus (in Luke 13:14–17) criticized hypocritical Pharisees whose actions clashed with their words. The Bible says much to support conscientiousness and impartiality. And practically every religion endorses GR, with many religions featuring GR as a summary of how to live.[57]

Virtue ethics. Ethics should emphasize *good character traits* instead of commands or ought-judgments. Formal ethics would be perfect if it were formulated in terms of character traits, like being consistent, impartial, conscientious, and fair-minded. So it would be better to formulate GR as: "A *fair-minded* person would habitually not treat another in a given way unless she was willing that if she were in the same situation then she be treated that

[56] Bernard Rimland, "The altruism paradox," *Psychological Reports* 51 (1982), pp. 221–2.

[57] For more on how GR connects to religion, see my *Ethics: A Contemporary Introduction*, pp. 154–61, *Ethics and the Golden Rule*, pp. 34–67 and 76–107, and *Ethics and Religion*.

way." And of course we'd add that being fair-minded is an important part of being a *good person*.

Hypothetical imperatives. Finally, we could express formal ethical principles as hypothetical imperatives about what we need to do to satisfy a certain ideal of moral consistency. Then GR would say: "To satisfy an ideal of *moral consistency*, don't treat another in a given way unless you're willing that if you were in the same situation then you be treated that way." This approach is neutral about metaethical frameworks.

So thinkers of diverse metaethical views could accept my approach to *formal ethics* – which clarifies and systematizes the idea that we should live in a way that's *consistent (including GR-consistent), informed, and imaginative*.

Harry J. Gensler