Value Theory

PHIL 640
Preliminary matters

- Fill out cards with basic info for my records.

- Ignore items intended for ug’s, and instead just list/describe prior work in ethics or other relevant areas, along with an indication of your own central interests. Let me know if you’ve read Joyce before.

- Registrants other than philosophy grad students need permission to take the course.

- Can someone be in charge of reminding us at 4:45 or so that it’s time to take a break?
Joyce, *The Myth of Morality*: an anti-realist argument, to be used critically, as a way of organizing diverse readings representing various metaethical views.

Darwall, Gibbard, Railton [= DGR] (eds.), *Moral Discourse and Practice*: a well-known anthology bringing together some of the work that Joyce discusses, from the 20th-century literature on metaethics.

Also, some supplementary readings online, e.g. recommended background readings from Hume’s *Treatise*. 
Requirements

- Midterm, final exams (take-home; 1-2 essay questions)

- At least eight comments on the weekly readings, initially expository, to be posted on Blackboard’s Discussion Board by Wednesday morning

- Presentations on some of our later readings: basically just leading discussion, ideally in a group

- Only the midterm and final are assigned letter grades, or a percentage of the final grade. Other work serves to boost grades from that initial basis (up to one full grade).
First assignments

- We’ll begin by reading Joyce’s Preface, but I’ll initially focus on providing some historical context. I’ve posted selections from Hume as recommended background reading; you might also read Moore in DGR.

- For the following week I’ve assigned Joyce’s discussion of noncognitivist anti-realism, along with the DGR excerpt from Stevenson as a sample. Some of Stevenson’s concepts are derived from Moore, and his reading of Hume needs correction.

- For the weekly writing assignment, you can make any comment (or raise any question) you like – or wait till we’ve read a bit more – but exegesis is recommended, particularly once we get past Joyce’s Preface. Don’t do an overview but instead focus on a particular passage, claim, or question.
Some basic contrasts

- Normative ethics
- Metaethics
  - Moral realism (e.g., intuitionism)
  - Anti-realism
    - Noncognitivism (e.g., emotivism)
    - Error theory
Humean background

- Hume is the first pre-20th-century figure who focused distinctly on metaethics as the basis for his ethical theory, though his place on the chart is disputed.
  - Many noncognitivists represent him as an early proponent of their view – though, oddly, Stevenson interprets him as holding (“in effect”) that moral judgments just describe majority approval, not necessarily the speaker’s.
  - On the other hand, Mackie claims Hume as a forerunner of error theory, the cognitivist anti-realist view that Joyce will be defending.

- Some contemporary philosophers question whether Hume is really such a skeptic about morality. He sees moral properties as not really located in objects, but rather in ourselves, by analogy to Lockean secondary qualities, but they’re still independent of individual minds.
Hume’s metaethical starting points

- Hume denies that reason can be the source of moral judgment. Its role is limited to discovering truth or falsity, i.e. facts.

- Morality motivates, and for motivation we need a passion (broadly construed to include dispositions such as benevolence, as “calm passions”).

- Neither a passion nor an action can “represent” anything – or be said to be true or false, reasonable or unreasonable, except derivatively, on the basis of accompanying factual judgments.

- So there’s a gap between “is” and “ought” – to be filled in by facts about human moral sentiments.
Hume gives a complex account of the development of moral sentiments, which he sees as primarily directed toward agents and their motives, i.e. he’s a virtue ethicist in normative terms (though a forerunner of utilitarianism).

His first stage turns on sympathy with others harmed or benefited by an agent:

- The initial basis for a moral sentiment simply amounts to the pleasure/pain we feel at benefit/harm to ourselves. By associating these impressions with an idea of the agent who gives rise to them, we form passions of love or hate, pride or humility.

- These and other feelings then get extended via sympathy to others benefited/harmed by someone. However, since sympathetic passions vary with our standpoint of observation – we react more to agents whose behavior affects others closer to ourselves – we don’t yet have moral sentiments.
Correcting sympathetic passions

- A second stage, of correction, is needed to allow for consistent uses of moral language across different standpoints of observation, as required by the communicative use of moral judgment.

- The corrected passion reflects a “general” or “common” point of view, the standpoint of an ideal (= ideally situated) observer or spectator, understood as someone in the circle of those affected by the agent’s actions.

- It’s only after this correction that we can be said to have moral sentiments, so they needn’t reflect our immediate reactions. Sometimes they even rest solely on supplying a term, without any corresponding feeling, for the sake of consistency with similar cases.
Kant offered a contrasting ethics of duty that didn’t depend on contingent human nature, but instead was supposed to hold necessarily for all rational beings who don’t automatically act in accordance with reason.

The central concept was that of a categorical imperative = an unconditional “ought” that gives reasons to anyone, regardless of his desires or other ends.

When twentieth-century social science made people more aware of cultural differences, Humean reliance on a constant “human nature” came to seem problematic, and many philosophers thought of the Kantian approach as embodying the alternative. Even skeptics like Foot and Williams understood morality in essentially Kantian terms, as we’ll see, and Joyce’s argument follows suit.
Many current non-specialists in ethics (e.g. writing about evolutionary ethics) conflate Hume’s “is”/”ought” gap with the “naturalistic fallacy,” derived from 20th-century moral philosopher G.E. Moore.

What Moore had in mind concerned definition: an inference from a general characterization of the properties basic to ethics – he focused on “good”-- to a definition in terms of naturalistic or other factual properties (e.g., pleasure, as suggested by Mill).

Moore comes up in Joyce much later, but he essentially kicked off 20th-century metaethics with an extreme version of moral realism, claiming that “good” refers to an indefinable property of objects, known by intuition.
Moore’s “open question” argument

Moore’s central argument for nonnaturalism is taken for granted by Stevenson and most other 20th-century authors, so here’s a sketch:

1. For any natural property N, suppose good =\text{df} N.

2. Then it shouldn’t even make sense (be an "open question") to ask whether something you recognize as N is good. It would be tantamount to questioning a tautology, asking: “Is this good thing good?”

3. But for any N, the question whether an N is good makes sense (is intelligible, or "open," even if we think the answer is “yes”).

4. So for any natural property N, good ≠\text{df} N.
The move to anti-realism

- Ross and others followed Moore in defending a view known as “intuitionism,” extending to “ought” as well as “good” (whereas Moore initially had held a version of utilitarianism [consequentialism], with terms for right action defined in terms of “good”).

- Besides its mysterious epistemological basis, intuitionism faced problems in explaining widespread moral disagreement – particularly with increasing awareness of other cultures.

- At the same time, general philosophic authors who embraced logical positivism, such as A. J. Ayer, could find no room in their epistemologies for moral knowledge. Later, linguists began to emphasize other possible functions of language besides describing putative facts.
Stevenson’s emotivism

- In a view that allows for descriptive as well as emotive meaning, Stevenson provides a better foundation for emotivism, the view outlined by Ayer in primitive (“Boo/Hurrah!”) form.

- Whereas Ayer had interpreted ethical statements simply as descriptively meaningless expressions of the speaker’s emotions, Stevenson assigns them a further, emotive *form* of meaning that doesn’t depend on describing distinctively moral facts.

- Instead of describing a natural or non-natural property of objects, “good” adds emotive meaning to the description of the speaker’s “interests” (= emotions and other pro/con-attitudes) – thereby leaving room for Moore’s “open question.”
Stevenson’s necessary conditions for a definition of “good”

1. makes sense of disagreement
   - vs. Hobbes’ subjectivist interest theory, which makes disagreement impossible, since we’re each just reporting our own desires
   - Stevenson substitutes “disagreement in interest” = opposing inclinations

2. possesses a certain “magnetism” (= motivational force; cf. p. 73)
   - vs. Hume’s [alleged] interest theory reporting majority approval
   - cf. contemporary [judgment-]internalism (and Hume’s *Treatise* III.I.II)

3. isn’t confirmable solely through scientific method
   - vs. all previous (i.e. descriptivist) interest theories
   - essentially a version of Moore’s “open question” argument against naturalism
Utterances with descriptive meaning can sometimes be used “dynamically,” to get people to do or feel something -- e.g. “It’s hot in here,” as used to get someone to open the window.

Emotive meaning amounts to the tendency a word has, by virtue of its history of dynamic use, to produce and/or result from affective responses.

A statement like “x is good” describes our interests but also, and more importantly, expresses them. in a way designed to subtly influence others to share them (e.g. via emotional contagion).
Stevenson offers two attempts at defining “x is (non-morally) good,” though he concludes that neither is adequate:

- “We like x” [used dynamically].
- “I like x; do so as well.”

These give a rough description of what we mean, but not a definition, since they lack the requisite subtle influence and hence can’t serve as “relevant” [= adequate] substitutes for “good.”

Stevenson ultimately interprets the question whether something is good that his inquiry was meant to clarify as a request for influence – in the first instance, by citing reasons, but once we agree about the facts, by non-rational means of persuasion.
Standard objections

- It certainly looks as though we *think* we’re describing something beyond our own reactions when we use moral terms.

- Also, Stevenson’s view limits the role of rational argument in ethics to clearing up disagreement about nonmoral facts, but can’t we argue rationally about value? A view that doesn’t allow for that seems unsatisfying, at any rate to philosophers.

- A more specific worry about rational moral argument is known as the Frege-Geach problem: whether and how we can make univocal sense of moral language in non-assertoric contexts (e.g., “if”-clauses, negation), of the sort essential to logical argument.
Joyce on defective discourses

- Joyce holds that moral discourse amounts to a faulty framework of the sort exemplified by defective concepts that are based on “non-negotiable” untrue propositions, such as
  - *phlogiston* (stuff stored in bodies and released in combustion, primary component of soot), and
  - *tapu* (contagious pollution).

- Non-negotiability is determined by a “translation test” (pp. 4f.) that Joyce will use his own ensuing argument: whether users of the discourse would be willing to take the term in question as translating a term in another language that isn’t held to satisfy the propositions.
An “error theory” holds that statements in a certain discourse are used to make assertions but typically fail to state truths.

- An argument for an error theory has two parts: conceptual (identifying non-negotiable elements of the meaning of a term) and ontological/substantive (determining that nothing satisfies them).

- An error theory is often thought to maintain that statements in the discourse are typically false, but Joyce adopts an interpretation of them as neither true nor false to accommodate Strawson’s account of non-referring expressions.

- Noncognitivism, the contrasting anti-realist view, is often characterized as denying that moral statements have truth values, but Joyce reformulates it as holding that moral judgments are typically used, not to assert something, but rather to express approval or a disguised command.
Error theory clearly fits *phlogiston*, but is noncognitivism more plausible for *tapu* (and moral discourse generally)? Not unless we take it as meaning something other than users of the discourse seem to think it means (according to the earmarks of assertoric discourse outlined by Glassen. p. 13).

Joyce treats Stevenson’s version of noncognitivism as pragmatic, in contrast to Ayer’s less plausible semantic version: the primary *use*, rather than the meaning, of moral language is non-assertoric.

He apparently thinks that Stevenson’s inclusion of descriptive meaning gets him by the Frege/Geach problem (see pp. 10-12 [but do we argue about our feelings?]).

Instead, he objects that we need to retain the appearance of assertion in order to manipulate others, by seeming to provide them with a reason. So on a Stevensonian account the influence of moral statements would seem to depend on either belief or feigning belief (see pp. 14f.), i.e. asserting something objective.
Fair against noncognitivism?

[Does Stevensonian magnetism require assertoric meaning?

- Can’t Stevenson also allow for persuasion without the pretense of belief, e.g. emotional contagion and other forms of social influence? Cf. also expression of second-order emotions (as in Blackburn), e.g. to shame a hard-hearted opponent of “the dole.”

- Cf. also, cf. Gibbard’s expressivism, which understands moral utterances as expressing a psychological state of norm-acceptance – rather than either belief or emotion (though the norm in question is a norm for assessing certain emotions)? We move toward consensus through pressure toward consistency in group discussion, with the function of promoting social coordination.]
Mackie’s cognitivist alternative

- Mackie accepts noncognitivism as a *revision* of our ordinary moral judgments, applicable to his own claims about normative ethics in the rest of the book from which the DGR excerpt is taken.

- But on his account ordinary moral judgments all involve a false – and even impossible – belief in objective moral values.

- He initially describes his metaethical position as a second-order version of skepticism, or a variant of subjectivism. But because of the confusing array of distinct positions coming under those terms, it’s generally referred to as (an) “error theory” (cf. p. 94).
Shifts in focus

Note that Mackie, in contrast to Stevenson, focuses primarily on characterizing moral requirements, rather than defining “good.”

Also, Moore’s nonnaturalism and Rossian intuitionism were by then out of favor, and Mackie’s characterization of recent swings between metaethical views is limited to noncognitivism and naturalism. What had intervened were

- a noncognitivist alternative to emotivism with Kantian elements: Hare’s “universal prescriptivism” (with moral judgments taken as expressing commands applicable to all relevantly similar agents) and

- variants of virtue ethics and utilitarianism with a basis in facts about human nature.
Mackie rejects both noncognitivism and naturalism as incapable of capturing the *authority* of ethics, understood in terms of the Kantian notion of categorical imperatives.

In Kant’s terms, moral requirements are necessarily binding, or “inescapable,” applying to all rational beings, whatever their ends.

What Mackie finds objectionable is the “objective, intrinsic prescriptivity” of moral requirements: their force for motivation, combined with their independence of our actual motivation.

[Nonmoral good and similar evaluative concepts lack this aspect and hence we can count them as objective without a problem.]
Mackie’s arguments

- **from relativity**: Cultural variation is better explained as reflecting divergent ways of life than in terms of something objective viewed from limited perceptual/evidential standpoints.

- **from queerness**: The relevant property, “objective prescriptivity,” would be
  - bizarre [or even incoherent?],
  - inexplicably related to natural properties, and
  - unknowable, except through a mysterious faculty of moral intuition.

- **from objectification**: Talk of objective moral properties can be explained [more parsimoniously] as a projection of social demands onto the world. in order to make them seem authoritative.
Mackie’s central argument, from queerness, seems to rest partly on combining objectivity with something like Stevensonian “magnetism” = motivational force.

- Joyce will interpret “prescriptivity” as normative rather than motivational: “to-be-doneness” (for acts deemed right) and “to-be-pursuedness” (for ends deemed good) involve reason-giving force.

- But as he acknowledges, at least part of what Mackie himself represents as bizarre is the thought that something independent of our motives could make us act, in the manner of Plato’s Form of the Good (see pp. 96f.).

As with Moore’s “open question” argument, even later authors holding contrasting positions tried to accommodate the “(judgment-)internalist” element of noncognitivism, i.e. motivational force as internal to the meaning of a moral judgment. Mackie seems essentially to be claiming that this rules out mind-independence.
Ethics without objectivity?

- Mackie apparently thinks we can just go on and do ethics (at any rate, utilitarian ethics) after rejecting objective moral values, since he has no problem with nonmoral values, such as rationality as applied to social ends.

- But could this sort of noncognitivist revision of ethics have the same kind of influence on others, unless it relies on their tendency to misinterpret it as objectively prescriptive?

- Joyce’s “fictionalism” will attempt to provide an alternative: a way of securing the social benefits of belief in objectivity, though we realize it’s just a pretense and aren’t relying on others’ being misled.
Joyce divides Mackie’s argument for error theory into two basic steps that he’ll retain:

- **Conceptual**: Moral discourse is committed to thesis $T$.
- **Ontological/substantive**: $T$ is false.

For Mackie, $T$ amounts to the claim that moral properties are “objectively prescriptive”: the universe issues prescriptions, in Joyce’s terms.

Joyce goes on to examine two ways of understanding $T$’s reference to prescriptivity: as a thesis about motivation (in the remainder of ch. 1), which he ultimately rejects, and as a thesis about the *reasons* provided by moral discourse (ch. 2).
Brink’s argument against Mackie depends on interpreting Mackie’s T as ascribing (defeasible) motivation to anyone who sincerely holds a moral judgment; see MI, p. 18 [= motivational internalism, aka judgment-internalism].

Though objections to MI have centered on the possibility of a totally amoral [= morally unmotivated] agent, the moral claims of such an agent could be said just to be reports of others’ opinions. Joyce instead argues for the possibility of a “purely evil” agent whose actions are explained by her sincere moral judgments.

His example is Sade’s Eugenie, who takes her moral judgments (sometimes in disagreement with others’ judgments) as reasons to do what they forbid. Eugenie isn’t weak-willed, so she’s also a counterexample to Smith’s reformulation of MI as MI* (p. 22), which assumes rationality.

However, it’s not so easy to show that moral discourse is non-negotiably committed to MI/MI*, as indicated by uncertainty about how we’d translate aliens’ use of “schmood,” as a term lacking motivational force (pp. 26f.).
In setting up his alternative interpretation of Mackie, Joyce modifies Foot’s early work on categorical vs. hypothetical imperatives.

- Foot herself starts out treating oughts derived from etiquette, club rules, and the like as “non-hypothetical” insofar as they wouldn’t be withdrawn [by those who accept the norms they presuppose] in application to agents who are indifferent to the norms.

- But she shifts to accommodate a Kantian interpretation of them as hypothetical insofar as they need support from something further to have any reason-giving force (p. 315, second column).

Her essential claim is that, however we classify them, their distinction from moral oughts can’t be maintained. Kantian talk of “inescapability,” and the sense in which we “have to” act morally, corresponds to nothing more than a feeling of compulsion.
Foot’s proposed alternative is a version of virtue ethics – one of the naturalistic approaches that Mackie claims is inadequate to capture the authority of morality.

- The virtuous person is essentially just an agent who’s characteristically motivated by certain ends, e.g. justice or charity.

- What Foot later “recants” (in a note tacked on to the article) is just her earlier assumption that [practical] rationality doesn’t require moral ends, or concern for the good of others. She attributes this assumption to her earlier acceptance of a Humean view of practical reasons, as dependent on agents’ contingent ends.

Joyce will move back to the Humean view of reasons, drawing from Williams’ defense of “reasons-internalism,” while accepting a Kantian picture of morality as requiring categoricity.
Basic normative alternatives

- Normative ethics
  - Virtue ethics (e.g., Aristotle, Hume)
  - Deontological ethics (e.g. Kant)
- Metaethics
  - Duty ethics
  - Consequentialism (e.g. utilitarianism)
Joyce on inescapability

- Joyce thinks Mackie’s argument from queerness won’t work for motivational force, but instead must rest on the “bindingness” or inescapability of morality: that it has jurisdiction even over agents who don’t make moral judgments and have no desires that are furthered by behaving morally (e.g. Plato’s invisible Gyges).

- He counts Foot’s nonmoral institutional oughts as categorical but weak, since they’re “non-evaporable,” applying even to agents who have no interest in participating in the institution, e.g. the unwilling gladiator Celadus, but fail to supply such an agent with reasons.

- What distinguishes moral oughts is supposed to be the fact that they give everyone reasons – though Foot’s article dismissed this Kantian claim and instead held that moral oughts could be hypothetical, i.e. applicable in virtue of agents’ shared desires and interests.
Joyce contrasts Mackie’s view with Kant’s and Foot’s and replaces MI with the inescapability of moral *reasons*. According to Mackie [practical] “ought” *means* “has a reason” (cf. MP = “Mackie’s Platitude,” p. 38).

But for an agent like Celadus, who stands outside an institution or system of norms, the reasons it offers to insiders (and that they can correctly ascribe to him as applications of the rules) don’t amount to “his” reasons [= reasons he “has” = reasons he can’t legitimately ignore].

Joyce distinguishes institutional from “real” reasons and supports Mackie’s claim with the case of an agent (p. 41) who’s completely unaware of some arbitrary scoring system that people he doesn’t know are using to assess him.
<table>
<thead>
<tr>
<th>Joyce on categorical imperatives</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>weak [= ~atc?]</th>
<th>strong [= atc?]</th>
</tr>
</thead>
<tbody>
<tr>
<td>= institutional</td>
<td>= binding, moral</td>
</tr>
</tbody>
</table>

- give ("real") reasons only to insiders, who can apply them to [certain] outsiders
- give ("real") reasons to everyone, regardless of interests or desires
Joyce’s argument, as sketched in basic form on p. 42, turns on the claim that it makes no sense to ascribe reasons to an agent that are independent of his desires or interests.

Against Carnap’s dismissal of “external” questions about the adequacy of a given framework as pseudo-questions, Joyce notes that the moral framework is nested in the wider framework of practical rationality.

But practical rationality isn’t itself an institution we can step out of. In raising questions about it, we’d be asking for further reasons, so the questions would be internal to it, after all (p. 47).

Humean instrumentalism takes promotion of the agent’s ends as the criterion of practical rationality, but Joyce’s next chapter will defend a non-Humean version of instrumentalism.
Williams on internal reasons

- Williams understands “internal” reasons (roughly) as reasons the agent has some motive for satisfying. He ultimately rejects the contrasting category, of “external” reasons.

- A simple “sub-Humean” model for the internal interpretation of reason-statements (p. 363) would make out the actions in question just as satisfying the agent’s desires.

- Instead, to allow for error about one’s reasons (cf. his gin/petrol example) and avoid restriction to pre-given desires, Williams characterizes internal reasons in terms of four propositions (see pp. 364f.) that make reasons dependent on the agent’s “subjective motivational set” (S) while ruling out dependence on false belief.
Against external reasons

- A nonmoral example of the contrasting category is provided by reasons relative to *needs* – e.g. to take a certain medicine to preserve one’s health – that fail to connect with a particular agent’s desires. But we’d probably make sense of this by filling in desires besides those the agent acknowledges.

- For another example of a putative external reason, consider Owen Wingrave’s father’s insistence that the family’s military tradition provides a reason for Owen to join the army, even though Owen has no supporting motive for joining.

- Williams rejects this on the ground that a reason has to be capable of explaining motivation. It could do so only if the agent came to believe it was a reason, but then the disposition to act on it would have become an element of S, so the agent *would* have an internal reason, after all.
Application to moral reasons

- Williams’s argument is meant to debunk Kantian attributions of moral reasons even to those with no supporting motivation – though he avoids comment on categorical imperatives or the relation of reasons to oughts (pp. 366f.)

- He allows that negative virtue-ethical terms (e.g. “selfish”) might be applied truthfully to agents like Owen who act in ways one thinks they should not.

- But he assumes that external reason statements are meant to convey a charge of irrationality for failure to act, and so they’re false. The charge amounts to mere “bluff” – and is unnecessary to justify following rules in “free rider” and similar cases, where doing so doesn’t serve our individual interests.
In ch. 3 Joyce sets out to argue against those who see practical rationality as yielding “real” moral reasons, independent of the agent’s ends. Instead, he defends a non-Humean form of instrumentalism (= reasons as means to an agent’s ends) for normative practical reasons.

He first allows for alternative reasons in cases like Williams’s gin/petrol case by distinguishing Williams’s objective reasons (to do what actually furthers one’s ends, on Joyce’s instrumentalist reading) from subjective reasons (those the agent is justified in believing to be objective reasons).

Rationality amounts to being guided by one’s subjective reasons, since irrationality entails arriving at one’s beliefs or actions in a faulty manner, for which one is responsible and deserves censure.
Vs. Humean instrumentalism

- For Joyce’s purposes, “ends” can be construed vaguely to cover both desires and interests (e.g. smoker Wanda’s interest in not smoking), though Hume restricts it to the agent’s present desires.

- Hume denies the possibility of practical rationality or irrationality and assumes that we always act on our [strongest] present desire.

- However, Joyce notes that intuitively it does seem that we can have and act on irrational desires. Rage and other emotional states can keep us from desiring what we think are the best means to our ends.

- Joyce’s central argument in ch. 3 will sketch an instrumental alternative to Hume that makes out our desires or ends as themselves subject to rational appraisal.
Some Humeans try to defend morality on the basis of hypothetical imperatives by trying to identify ends all humans have, such as self-interest and/or sympathy.

- But self-interest won’t always support morality in particular cases.

- Moreover, as Hume recognized, sympathy isn’t something we always feel, or feel for all fellow humans, as opposed to our inner circle.

Foot instead makes out morality based on hypothetical imperatives as escapable by those who don’t share the relevant ends. However, that doesn’t fit ordinary moral discourse.

[Cf. his Nazi case, p. 62. Does condemnation imply an attribution of reasons to someone whose desires don’t support morality?]
Smith’s non-Humean theory

- Michael Smith argues from a non-Humean theory of reasons to a version of moral realism, but his argument rests on many questionable claims, e.g. MI* from ch. 1 and the claim that we can expect a rational agent to act morally.

- However, Joyce agrees that we need a non-Humean theory to handle weakness of will – which Smith makes out as acting contrary to our values (vs. our strongest desire, as in Hume). Values amount to the desires that survive rational deliberation about our initial stock of desires (see pp. 71f.).

- Joyce also follows Smith (among others) in distinguishing normative (justifying) reasons from motivating (explanatory) reasons.

- He takes rational reasons as based on desires that would survive rational deliberation – vs. irrational reasons (capturing the remaining Humean reasons; cf. Molly’s case, pp. 72f. and p. 74).
Smith’s own notion of rationality (like Brandt’s, discussed earlier in the chapter) implausibly requires possession of all true beliefs.

However, to test Smith’s claim that subjectively rational reasons [= values] yield reasons with the content of ordinary morality, Joyce has us consider an idealized version S+ of agent S who gives S rational advice (cf. Railton).

Whether or not one attributes all relevant true beliefs to S+, it looks unlikely that there would be much convergence in different agents’ idealized desires, as needed to support morality. This point will be argued in chs. 4-5.
Joyce argued in ch. 2 that practical rationality is the source of inescapable reasons and then went on to defend a non-Humean instrumentalist account of practical reasons in ch. 3.

Ch. 4 is aimed at eliminating two opposed rivals to moral error theory: Smith’s rationalism and Harman’s relativism.

[Note that Harman’s relativism is a revisionist view about the nature of morality, whereas Joyce’s defense of “relativism” in the earlier part in the chapter pertains to practical reasons and their implications for morality.]
The rationalist’s dilemma

- A rationalist like Smith (= someone who views a moral violation as irrational) faces a choice between
  
  1. letting rational agents be alienated from [some of] their acknowledged normative reasons, and

  2. making such reasons relative to a particular agent’s desires, which would not yield an adequate basis for moral reasons.

- Joyce thinks Smith manages to avoid 1 by basing morality on an ideally rational version of the agent (vs. Firth’s Ideal Observer).

- But avoiding 2 requires the assumption that full information and careful reflection would eliminate differences in rational agents’ desires.
To ground categorical moral imperatives on rationality, according to Joyce, Smith has to deny that rational agents could ever be indifferent to what they recognize as normative practical reasons.

[Is this so obviously impossible, though? Both Smith and Joyce assume a Williams-style basis of reasons in an agent’s own desires, but there are alternatives phrased in terms of promotion of value (e.g., Raz).]

Joyce undermines Smith’s defense of convergence (see pp. 88ff.) by exhibiting alternative explanations for cases where agreement is reached, e.g. by mutual social influence plus the need for cooperation.
Harman comes up initially as the source of a case that counts against Smith’s insistence on rational convergence: a mobster employed by Murder, Inc., who stands to gain financially by murdering a non-family member, Bernard Ortcutt.

Harman’s point is that we wouldn’t be providing the mobster with a reason by appealing to moral considerations, though he isn’t in any way irrational. [Cf. Williams on Owen Wingrave.]

Joyce takes this to illustrate his claim that instrumental reasons are agent-relative. But since moral reasons are conceptually non-relative, Smith must be wrong in claiming they’re derivable from instrumental reasons. So moral rationalism is false.
Vs. Harman’s moral relativism

- Harman’s own relativistic moral view represents another alternative to error theory insofar as it’s designed to save ordinary moral discourse – by making out moral judgments as elliptical for the agent-relative judgments that are derivable from instrumental reasons.

- But Harman’s analogy to the way we save pre-Einsteinian judgments of motion in light of relativity isn’t apt, Joyce says, because here we’re not accommodating a new discovery. Moral relativism has long been recognized and denied by ordinary moral discourse.

- We’ll be abolishing moral discourse, on the model of phlogiston or witch discourse, if we reject absolutism.
Using Williams’s argument

- Williams takes all reasons as internal, meaning that they would motivate a fully informed version of the agent who deliberated correctly.

- Joyce limits his argument against external reasons to normative (as opposed to institutional) reasons, using Molly’s case from ch. 4 to illustrate the various types of reason (among other things, distinguishing objective from Williams-style reasons; see pp. 107f.).

- The argument rests on both the Humean assumption that motivation requires a desire (not just a belief) and an understanding of reasons as potentially explaining action.

- Joyce goes on to defend the argument against objections (see next slide for a sketch, with issues for discussion on the following slide).
Sketch of replies to objections

- Millgram’s example of a reason – to (see a therapist to) become more sensitive – that Archie doesn’t care about but would find satisfying to act on could explain action by Archie’s becoming increasingly like Archie+, deliberating better on fuller information. But then the reason would be internal, after all.

- Hampton confuses the (false) claim that we always have an external reason to deliberate correctly with the claim that we always ought to act in accordance with correct deliberation (i.e., take S+’s advice). In any case, her external reason wouldn’t yield reasons with moral content.

- Korsgaard’s insistence on agent-neutral reasons confuses agent-relative with private reasons and misapplies Wittgenstein’s “private language” argument. Valuing his own humanity needn’t supply Murder, Inc.’s “Al” with a reason to value it in others, unless it gains support from his desires.
Some issues for discussion

- Doesn’t Archie’s case illustrate how an agent may have a reason just in virtue of his current beliefs – if knowledge of his future desires is enough to yield a reason to do what’s needed now in order to bring about their satisfaction?

- Does Joyce’s rejection of reasons with conditional content, in response a version of Hampton’s argument, as yielding “too many reasons” jibe with his willingness to allow all sorts of reasons in his earlier treatment of Williams? The issue here seems to be outlandish content, but what about trivial logical conclusions from recognized reasons, e.g. by disjunction?

- Might Korsgaard be assuming that a rational person couldn’t just value one instance of humanity, namely his own, but rather would be committed to valuing humanity in general? To pick out one instance might seem arbitrary. But is it?
Joyce’s qualified conclusion

- That completes Joyce’s argument for moral error theory – which he now limits to establishing that it’s *reasonable to believe*.
  - He notes that he doesn’t take Williams’s argument to rule out all external reasons, since institutional reasons may be external.
  - Instead, the argument cuts against *normative* external reasons, reasons said to transcend all institutions, as in the case of moral reasons (see p. 134).

- The burden is on rationalists to show how their claim that morality possesses a distinctive kind of authority is anything more than bluff.
An evolutionary rationale

- Joyce explains the error in error theory by appealing to the greater reproductive fitness achieved by adding to natural sympathy for kin an innate sense of requirement.

- Kin-helping develops into reciprocal altruism by way of hostility toward exploiters, making cooperative tendencies more reliable insofar as they’re independent of either sympathy or fear of payback. (Cf. Stig’s choices, pp. 143f.)

- At the point where prototypical categorical imperatives emerge, Joyce begins to speak of moral appraisal and even some specific moral beliefs as innate (see p. 146).
Some authors have proposed an evolutionary “success” theory, using Hume’s analogy between moral properties and secondary qualities like color [as rendered compatible with realism by contemporary philosophers like McDowell].

However, a dispositional analysis of “rightness” in terms of the tendency to evoke approval isn’t sufficient to explain the sense of binding requirement, as entailing reasons independent of desire.

That we have an evolved disposition to promote community good (as Richards argues) doesn’t imply that we all have a desire to promote it on each particular occasion. But without that [or some other desire supporting it], what reason do we have?
Moore’s “naturalistic fallacy” is often misinterpreted by evolutionary theorists as a version of Hume’s “is/ought” argument applied to “good,” but instead it concerns the *definability* of the concept.

The fallacy amounts to inferring identity from predication [e.g., deriving the definition of good from the claim that a certain sort of thing, e.g. pleasure, is good]. But naturalism (defined on p. 154) needn’t be based on that confusion.

Even if it’s not based on a fallacy, though, naturalism doesn’t seem to yield an adequate account of moral discourse. The strategy of revising the meaning of a concept to fit its extension would also vindicate discourse about witches and the like (cf. p. 156).

Just as supernatural powers are essential features of witches, the whole point of moral discourse is its categorical force.
Joyce argues that the evolutionary explanation of the categorical aspect of morality as innate would undermine the justification for moral judgments, since it would imply that they’re insensitive to evidence (cf. analogy to paranoia, p. 159).

To establish that they’re false, though, we need to rely on Joyce’s arguments in earlier chapters. The usefulness of moral discourse (which Joyce will exploit in ensuing chapters) doesn’t mean that moral judgments are true. Cf. p. 168 for Joyce’s conclusion that they’re *probably* untrue, on the basis of appeal to parsimony.

Other moral terms besides “ought” and “wrong,” e.g. “good,” “depraved,” and terms for particular virtues, turn out to be connected to categorical imperatives (see pp. 175f.). [So the problem isn’t escaped by turning from Kant to virtue ethics, as favored by Williams and others.]
Railton offers a non-evolutionary naturalist view that he defends as a version of moral realism.

Moral judgments on his account don’t necessarily give reasons to all rational beings, though they’re still applicable to everyone. They’re also non-relativist, though based on an agent-relative notion of nonmoral value.

The aims of his argument are twofold (though often intertwined):

- to defend the possibility of reducing values (and ultimately social norms) to the natural properties on which they supervene, and
- to show that values (norms) serve to explain changes in desires (moral codes) via a feedback mechanism [contra the argument by Harman in DGR that moral judgments are explanatorily superfluous].
<table>
<thead>
<tr>
<th>SCIENCE</th>
<th>ETHICS</th>
</tr>
</thead>
</table>
| **theoretical posit**<br>○ protons<br>[needed for best explanation of:]
| **theoretical posit**<br>○ wrongness/moral facts<br>[not needed for best explanation of:]
| **observation**<br>○ “There goes a proton.”
| **observation**<br>○ “That’s wrong.”
| **circumstances**<br>○ sees vapor trail
| **circumstances**<br>○ sees kids burning cat |
General stages of Railton’s argument

1. Derivation of facts about *nonmoral value* for a particular agent (= his objective good, or his interests) from his actual desires (= his subjective good), as modified to reflect what an idealized version of himself would want him to want.

2. Extension to facts about *individual rationality* via interpretation of rational norms in terms of criteria for assessing an agent’s behavior by its success or failure in achieving his objective good.

3. Extension to facts about *social rationality*, understood in terms of the aggregate of individual objective goods, and amounting to moral norms.
Objectivizing subjective value

- Start with the *subjective interests* of A = what A happens to want (p. 142; cf. example of Lonnie, p. 143).
  - A’s desires amount to secondary qualities of A.
  - These supervene on primary qualities (of A’s constitution, circumstances, etc.) = the *reduction basis* for subjective interests.

- Then take A’s *objective* interests as what an idealized version of A, A+ (with full information and imaginative/cognitive capacities), would want A to want.
  - = what really is in A’s interests, whether or not he actually wants it
  - *reduction basis* = facts about A that A+ would take into account
Establishing value realism

- We then can take note of a \textit{wants-interests mechanism} (pp. 144f.): a feedback loop whereby largely unreflective trial and error, resulting in better satisfaction of A’s objective interests, modifies his wants to fit them better.
  - Since the mechanism can operate independently of belief, it has a genuine explanatory role [contra Harman’s argument from the availability of a better explanation of an agent’s beliefs in terms of his upbringing].
  - What’s explained by objective values includes changes amounting to conscious or unconscious \textit{learning}: the evolution of subjective in the direction of objective interests (cf. Lonnie).

- But so far we have only \textit{nonmoral} value.
Extension to normative realism

- Normative facts (e.g. the hypothetical oughts of instrumental rationality) are understood in terms of *criterial* explanation (p. 145ff.), as summing up a process that selects for achievement of some goal (what a roof is for; what a person wants, his objective interests, etc.).
  - So for individual rationality (= prudence; e.g. the claim that Lonnie ought to drink clear liquids when traveling) the goal is achievement of objective interests (= the agent’s good), replacing mere desire-satisfaction.
  - There’s a “selection” process influencing behavior that works via failure and consequent dissatisfaction but doesn’t necessarily register in beliefs about what’s rational [thus avoiding Harman’s appeal to a better explanation in terms of sources of belief.]

- The upshot amounts to *facts* about what individuals have reason to do, i.e. what would be rational for them (as required by realism about individual rationality).
Moral norms (pp. 149ff.) may then be understood in terms of *social rationality* (= what we’d be rationally advised to do under circumstances of full and vivid information, counting everyone’s interests equally).

Here the criterial goal amounts to satisfying everyone’s aggregated interests (to the extent possible). The wants/interests mechanism in this application operates by way of social dissatisfaction or unrest, as negative consequences of discounting the interests of a social subgroup. This need not rest on forming a *belief* that society is unjust [so that a better explanation in terms of other sources of moral belief, of the sort Harman’s argument relied on, isn’t available].

Railton takes his theory to be confirmed by three patterns in the evolution of moral norms: (1) generality (widening extension beyond society), (2) humanization (basis in effects on human interests), (3) patterns of variation (more uniformity under certain conditions).
Forgoing categorical reasons

- Morality on Railton’s account depends on an impartial, but a human, point of view – as opposed to Kant’s extension of moral oughts to all rational beings, or the classical utilitarian extension to all sentient beings.

- Morality also depends on a social point of view – and hence isn’t categorical, in the sense of giving reasons to all individual agents.

- The social point of view includes one’s own, but also is capable of conflicting with it. It yields general grounds for moral behavior, even if not personal reasons.

- Moral oughts still apply to everyone, though not as a matter of individual rationality.
A substitute for morality?

- Whether Railton’s view really rests on error theory depends on what we think is essential to morality. Do we really have to take everyone to have a reason to act morally – in Joyce’s instrumentalist sense of reason, as based on the agent’s actual ends?

- Isn’t it enough to have adequate reason *ourselves* to be justified in condemning someone who doesn’t have moral reasons (in Joyce’s sense)? Even if people ordinarily assume something further – or say something further to influence others – categoricity needn’t be taken as a necessary feature of moral discourse.

- Among philosophers there’s also a common assumption that it would be unfair to blame someone for failure to have a certain desire, since we can’t help whether we care about something. That’s what seems to be behind Joyce’s treatment of the Nazi case.
Joyce’s alternative

- After undermining ordinary moral discourse as he sees it, Joyce favors a cost/benefit analysis of morality, based on preference-satisfaction.

- He takes the instrumental justification of morality to rest on its role in bolstering self-control. However, believing the truth also has instrumental importance, so he argues that we can gain the benefits of moral discourse without believing in it.

- Instead, we should accept it as a fiction, confining questions about its truth to special critical contexts, including but not limited to philosophical contexts.
An instrumental justification

- The result of a cost/benefit analysis will be the instrumental justification of a policy, not of each and every moral action.

- It won’t provide an alternative basis for moral belief, i.e. evidential grounds for the content of moral judgments.

- What it justifies is the widespread adoption of a general attitude that Joyce will call “acceptance” rather than belief.

- Joyce also claims that the right moral attitudes amount to the most useful. [Cf. his discussion of the Tasmanian genocide, pp. 181ff.]
Acceptance without belief

- Vaihinger’s “as if” philosophy involved a form of fictive belief that applied to nonmoral concepts too. But it committed agents to both believing and disbelieving the same propositions.

- So Joyce substitutes “acceptance” – of propositions one has denied, and retains the tendency to deny, in more critical contexts (see pp. 192f.), and hence disbelieves.

- Critical contexts extend to ordinary agents, not just an intellectual elite, and are assumed to be contexts we can enter into voluntarily, so that acceptance of something we disbelieve doesn’t entail being self-deceived.
There’s an analogy to attitudes we take toward literary fictions, though morality lacks an author or canonical text. Cf. pp. 195f. for the Sherlock Holmes fan who gets pleasure from pretending the stories are real but isn’t self-deceived.

Make-believe involves entertaining *thoughts* that can engage our emotions without belief (p. 197).

Joyce makes out fictive moral discourse as nonassertoric – so that, if adopted, it would yield noncognitivism. But the need for surrounding linguistic conventions means that it has to be adopted by a *group*.
In his final chapter Joyce argues that moral discourse would indeed still benefit those who withhold belief from it.

It can deter Hobbes’s “Foole” or Hume’s “sensible knave” from defecting from cooperative schemes by putting an internalized tax on defection.

Moral discourse serves as a bulwark against temptation, since categorical imperatives silence calculation of what an agent has to gain by failing to cooperate in a particular instance.

Belief in morality might be more effective, but false belief has costs, and a habit of entertaining thoughts to the same effect can yield many of the same benefits.
However, Joyce acknowledges that moral discourse, on his construal as offering practical advice, would not deter unusually situated agents like Plato’s Gyges, who has a ring that makes him invisible.

Yet morality still applies to such agents. Indeed, the appropriateness of moral condemnation of them was part of the Joyce’s argument for error theory. But our claim that they deserve punishment has to be taken as just a useful fiction.

What keeps the rest of us from acting immorally when we can get away with it is the fact that our acceptance of morality is a “precommitment” – analogous to Ulysses’s having himself tied to the mast – rather than a product of case-by-case calculation. We’ve formed habits of thought on the basis of childhood beliefs.