Utilitas (2019), **31**, 277–290
doi:10.1017/S0953820819000165

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Constructivism about Intertheoretic Comparisons

Stefan Riedener

University of Zurich
stefan.riedener@philos.uzh.ch

Many people think that if you're uncertain about which moral theory is correct, you ought to maximize the expected choice-worthiness of your actions. This idea presupposes that the strengths of our moral reasons are comparable across theories – for instance, that our reasons to create new people, according to total utilitarianism, can be stronger than our reasons to benefit an existing person, according to a person-affecting view. But how can we make sense of such comparisons? In this article, I introduce a constructivist account of intertheoretic comparisons. On this account, such comparisons don't hold independently of facts about morally uncertain agents. They're simply the result of an ideal deliberation in terms of certain epistemic norms about what you ought to do in light of your uncertainty. If I'm right, this account is metaphysically more parsimonious than some existing proposals, and yet has plausible and strong implications.

I. Introduction

What ought you to do if you're uncertain about which moral theory is correct? For example, suppose you think that either total utilitarianism (TU) or a person-affecting deontological theory (PAD) must be true, find PAD somewhat more plausible than TU, but are ultimately uncertain between the two. And suppose you can either bring into existence a massive population of people with lives well worth living (option *a*) or benefit an existing person slightly (option *b*). TU says you ought to choose option *a*, and PAD says you ought to choose *b*. What ought you to do?

A prominent idea is to extend standard decision theory to such cases. Under empirical uncertainty we sometimes ought to hedge for the possibility that our best assumptions may be false. We ought to buy a fire extinguisher, say, even if it's unlikely that our house will ever catch alight. More specifically, the standard theory of decision-making under empirical uncertainty is that we ought to maximize the expected value of our actions. So prima facie, something similar should be true in the moral case. It seems we sometimes ought to hedge for the possibility that our favourite moral theory may be false, or maximize the expected choice-worthiness of our actions.[1] In the case at

[1] See e.g. T. Lockhart, *Moral Uncertainty and its Consequences* (Oxford, 2000); A. Sepielli, 'Along an Imperfectly Lighted Path' (PhD dissertation, Rutgers University, 2010); W. MacAskill, 'Decision-Making under Normative Uncertainty' (PhD dissertation, University of Oxford, 2014); W. MacAskill and T. Ord, 'Why Maximize Expected Choice-Worthiness?', *Noûs* (forthcoming); S. Riedener, 'An Axiomatic Approach to Axiological Uncertainty', *Philosophical Studies* (forthcoming).

https://doi.org/10.1017/S0953820819000165 Published online by Cambridge University Press

hand, it seems you ought to bring about the massive population, even if you find PAD slightly more plausible than TU.

The most serious obstacle for this idea is the problem of intertheoretic comparisons. Hedging against the case of a fire may be reasonable because the *stakes* are much greater in that case. If your house stays safe it's only slightly worse to have bought an extinguisher than to have saved the money, whereas if your house burns at some point it's *much* better to have bought one. The same will be true in the moral case. In our example, hedging can be reasonable only if the stakes are greater for TU than for PAD – or as I'll put it, if

> the moral reason you have for doing *a* rather than *b*, according to TU, is stronger than the moral reason you have for doing *b* rather than *a*, according to PAD.

I'll call a statement of this kind an *intertheoretic reason-comparison*, or IRC. Many people are sceptical about such statements. Edward Gracely, for example, asks: 'is a small loss of utility as seen by a [person-affecting theory] more or less important under that theory than a large loss of utility (involving lives not created) under total utilitarianism?'[2] He says, 'I don't quite see how this question could be answered', as 'there is no abstract scale of "wrongness" outside of the rank provided *within* a theory',[3] and 'comparisons of rankings and weightings between different theories are essentially meaningless'.[4] Other people have expressed similar scepticism, and concluded that the idea of hedging under moral uncertainty is doomed.[5]

What precisely is the problem of intertheoretic comparisons? It's helpful to distinguish three worries that jointly form this problem. First, there's a question about the *meaning* (or criterion of identification) of IRCs: what *is it* for an IRC to hold? We might understand what it is for reasons to compare intra-theoretically in a certain way – e.g. what it is for your reason to create a massive population to be stronger than your reason to benefit an existing person slightly, according to TU. But it seems unclear what such comparisons across theories even amount to. Yet suppose we know what it would be for an IRC to hold. There's then a second, object-level question about the actual *IRC-facts*: do any IRCs hold; or if so, which of them do? For instance, is our reason not to betray a friend, according to PAD, as strong as our reason to bring into existence a happy person, according to TU – or a happy cat, or a small population of people? It seems unclear which such comparisons hold, or indeed whether any of them do. Yet suppose we know that some IRCs *do* hold, perhaps in some simple cases. There's then a third question about the *grounds* of these facts: what is it that grounds IRCs? It seems that moral theories themselves don't say how strong our reasons would be, if certain other theories were true. So one might wonder what the basis of IRCs can be.

[2] E. Gracely, 'On the Noncomparability of Judgments Made by Different Ethical Theories', *Metaphilosophy* 27 (1996), pp. 327–32, at 331.

[3] Gracely, 'Noncomparability', p. 331.

[4] Gracely, 'Noncomparability', p. 330.

[5] See e.g. J. Hudson, 'Subjectivization in Ethics', *American Philosophical Quarterly* 26 (1989), pp. 221–9; J. Broome, *Climate Matters: Ethics in a Warming World* (New York, 2012), p. 185; J. E. Gustafsson and T. O. Torpman, 'In Defence of My Favourite Theory', *Pacific Philosophical Quarterly* 95 (2014), pp. 159–74; I. Nissan-Rozen, 'Against Moral Hedging', *Economics and Philosophy* 31 (2015), pp. 349–69; B. Hedden, 'Does MITE Make Right? On Decision-Making under Normative Uncertainty', *Oxford Studies in Metaethics*, vol. 11, ed. R. Shafer-Landau (Oxford, 2016), pp. 102–28.

https://doi.org/10.1017/S0953820819000165 Published online by Cambridge University Press

The aim of this article is to sketch an answer to these questions – or more precisely, to outline the foundations of a proposal that's metaphysically more parsimonious than some existing accounts with respect to the meaning and grounds of IRCs, and yet delivers plausible and strong object-level results about which IRCs hold. At bottom, I'll defend a version of the 'ought-first' approach. That is, I'll suggest that for the above IRC to hold just is for it to be the case that if you find both TU and PAD equally likely (and these are the only theories you have credence in), you ought to do *a* rather than *b* in light of your uncertainty. This idea has been mentioned in a brief passage by Jacob Ross,[6] but it hasn't yet received the exploration it deserves. In particular, we don't yet have a compelling theory about what grounds these ought-facts, and about which ought-facts hold, on an ought-first approach. So I'll propose a version of this approach, on which ought-facts are grounded in epistemic norms. In other words, I'll propose a form of *constructivism about IRCs*. If I'm right, IRCs are not facts out there that hold independently of facts about morally uncertain agents. They hold in virtue of being the result of an ideally reasonable deliberation, in terms of certain epistemic norms, about what you ought to do in light of your uncertainty.

To explain and defend these ideas, I'll first discuss three existing accounts of IRCs (section II). The purpose of this is not to refute these proposals. I don't have any knock-down arguments against them. But the discussion will at least motivate a further exploration of the space of possible views, and show what's distinctive and attractive about the constructivist ought-first account that I'll then propose (sections III–V).

Three clarifications before I begin. First, I'll set aside any more fundamental worries about the meaningfulness or importance of moral uncertainty. Though this has been challenged,[7] I'll simply assume that talk of moral 'uncertainty' makes sense, and that there's a relevant 'ought' under such uncertainty.[8] Second, for simplicity, I'll just speak about what you ought to do 'in light of your uncertainty', or your moral credences or beliefs, rather than in light of your evidence, your evidential probabilities or the beliefs your evidence warrants. But for the purposes of this article, nothing hinges on this. My arguments could be made *mutatis mutandis* with respect to the uncertainty, or beliefs or credences, that your evidence warrants. Third, note that the idea of expected choice-worthiness maximization actually presupposes *cardinal* IRCs – i.e. statements of the form 'the moral reason you have for doing *a* rather than *b*, according to TU, is *n times as strong as* the moral reason you have for doing *b* rather than *a*, according to PAD'. This adds an important layer of complexity to the problem as I've stated it. But I'm interested in the more fundamental question about how *any* (even ordinal) IRCs can hold. And as I'll indicate, my proposal can readily account for cardinal IRCs. So for most of the article, the difference between ordinal and cardinal IRCs won't matter.

[6] J. Ross, 'Rejecting Ethical Deflationism', *Ethics* 116 (2006), pp. 742–68, at 763.

[7] See e.g. E. Harman, 'The Irrelevance of Moral Uncertainty', *Oxford Studies in Metaethics*, vol. 10, ed. R. Shafer-Landau (Oxford, 2015), pp. 53–79; B. Weatherson, 'Running Risks Morally', *Philosophical Studies* 167 (2014), pp. 141–63; Hedden, 'MITE'.

[8] For arguments, see e.g. C. Tarsney, 'Rationality and Moral Risk: A Moderate Defense of Hedging' (PhD dissertation, University of Maryland, 2017); A. Sepielli, 'What to Do When You Don't Know What to Do When You Don't Know What to Do …', *Noûs* 47 (2013), pp. 521–44.

https://doi.org/10.1017/S0953820819000165 Published online by Cambridge University Press

II. Three existing proposals

II.1. Structural accounts

There are already a number of proposals about IRCs. So it's worth starting with a brief discussion of these. Consider first 'structural accounts'. According to these proposals, IRCs are grounded in general principles of rationality about how to normalize moral theories for decision-making under moral uncertainty. And these principles take into account only *structural* features of the theories – i.e. features of the theories' (ordinal or cardinal) ranking of options in terms of how much reason you have to choose them.

Various principles of this kind have been proposed. For instance, Ted Lockhart suggested the 'Principle of Equity among Moral Theories', according to which, in every choice-situation, the reason to choose the best rather than the worst option should be considered equally strong according to all theories.[9] Andrew Sepielli discussed (but didn't endorse) a variation of this principle, according to which the reason to choose the best rather than the worst *conceivable* option should be considered equally strong according to all theories.[10] William MacAskill suggested that the *variance* of theories should be considered equal, where the variance of a theory is a measure of how moral choice-worthiness is spread out over different options – viz., the average of the squared differences in choice-worthiness from the mean choice-worthiness of options.[11] And infinitely many other structural proposals can be imagined beyond these.

What's nice about these accounts is that they're metaphysically parsimonious. They ground IRCs fully in principles of rationality, and don't assume that there's an antecedent fact of the matter about how theories compare. Certainly, all of the above proposals have their specific problems.[12] So it remains to be seen what the most plausible principle would be. However, these accounts also face a general problem. In so far as we have intuitions about how to make IRCs in certain simple cases, these intuitions are sensitive to the *content* of moral theories. For example, suppose you're certain that consequentialism is true and that pleasure has value, but uncertain whether beauty also has value, and that this is the only moral uncertainty you have. We can then describe you as being uncertain between two theories, a monist theory on which only pleasure has value and a pluralist theory on which pleasure and beauty have value. Intuitively, it seems reasonable to compare your two theories in such a way that the reasons to bring about pleasure are the same on both theories. After all, you're not uncertain about these reasons. You're only uncertain about the additional reasons of beauty. Purely structural accounts cannot capture this content-based intuition.

More specifically, the guiding idea of standard structural principles is that the moral stakes should somehow be considered equal according to all theories. But in so far as we have intuitions about IRCs, it seems that the stakes may be higher on some theories than on others, due to their content. For instance, it seems that if *both* pleasure and beauty have value, the moral stakes (overall, or in some choice-situations) are higher

[9] Lockhart, *Moral Uncertainty*, p. 84.

[10] A. Sepielli, 'Moral Uncertainty and the Principle of Equity among Moral Theories', *Philosophy and Phenomenological Research* 86 (2013), pp. 580–9, at 588.

[11] MacAskill, 'Decision-Making', p. 89.

[12] For instance, Lockhart's principle can require you knowingly to choose a course of action that's worse than some available alternative on every theory in which you have credence (see Sepielli, 'Moral Uncertainty'). Sepielli's proposal can't apply to the numerous theories on which there *are* no best and worst conceivable options (as he notes). And variance-normalization faces some technical challenges in order to be well-defined (see MacAskill 'Decision-Making', p. 104n. and p. 76).

https://doi.org/10.1017/S0953820819000165 Published online by Cambridge University Press

than if only pleasure has value. Again, purely structural accounts can't capture this intuition.[13] So the cost of their parsimony, it seems, is that they have implausible implications. Other things equal, we should prefer accounts on which IRCs are content-sensitive.

II.2. Metaphysical accounts

A range of accounts that *are* content-sensitive is what I'll call 'metaphysical accounts'. On these accounts, IRCs are *not* grounded in any facts about morally uncertain agents – i.e. in criteria of rationality for decision-making under uncertainty, or in epistemic principles, or actual beliefs of such agents. Rather, they're grounded in facts about values or reasons themselves, and are in this sense 'metaphysical facts' out there.

The most explicit such account has been defended by Christian Tarsney.[14] So let me consider his version. Tarsney starts precisely from the comparison between the monist pleasure-theory and the pluralist pleasure/beauty-theory I've just considered. To account for the intuitive IRCs between these theories, he suggests there are facts like

> *Independence:* the strength of our reasons to bring about pleasure is independent of whether we have non-derivative reasons to bring about beauty.[15]

Tarsney understands Independence as holding independently of any facts about morally uncertain agents. So as he understands it, Independence is a normative fact about reasons, quite like the fact that we have reasons to bring about pleasure (if this is a fact). It's simply a *counterfactual* normative fact, about how strong our reasons of pleasure would be if we had non-derivative reasons of beauty.[16] I'll call a statement of this kind a *reason-counterfactual*. If some such counterfactuals hold, then they straightforwardly ground IRCs. For example, if Independence holds, the reasons to bring about pleasure are the same on the monist and on the pluralist theory. So this proposal can straightforwardly account for our content-based intuitions.

However, this strategy also faces problems. To begin with, it doesn't seem to provide a sufficient story about what it *is* for IRCs to hold. At best, this question is now pushed back to reason-counterfactuals. And these are far from self-explanatory. Suppose that the pleasure-theory is correct. What should it *mean* (say) that if beauty had value, our reasons of pleasure would be half as strong as they actually are? Or even simpler, what would it mean that if beauty had value, our reasons of pleasure would be weaker than they now are? And what should it mean, for that matter, that if beauty had value, our reasons of pleasure would still be exactly as strong as they are? Intuitively, we don't understand these claims unless some further explication is given for them. So it seems that the sceptical challenge of explaining what IRCs amount to really still remains.

But suppose we do have a sufficient intuitive grasp, or some helpful explication, of statements like Independence. There are then still worries about the object-level facts and their grounds. It's controversial that the universe contains *any* mind-independent normative facts. But it seems quite an ontological burden to assume that it should

[13] The same has been argued by MacAskill, 'Decision-Making', p. 134.

[14] Tarsney, 'Rationality and Moral Risk'; C. Tarsney, 'Intertheoretic Value Comparison: A Modest Proposal', *Journal of Moral Philosophy* 15 (2018), pp. 324–44.

[15] See Tarsney, 'Rationality and Moral Risk', p. 312.

[16] See Tarsney, 'Rationality and Moral Risk', p. 338.

https://doi.org/10.1017/S0953820819000165 Published online by Cambridge University Press

contain such *counterfactuals*. Suppose again that the pleasure-theory is correct. Why should there be any fact of the matter about how the reasons implied by a *false* moral theory like the pluralist view compare to our actual reasons? Why should the fabric of the universe contain not just standard normative facts, but also counterfactuals about how strong our reasons *would* be if we had certain other reasons that in fact we don't have?

But suppose we grant that there are reason-counterfactuals like Independence – counterfactuals to the effect that the strength of certain reasons wouldn't change if there were *additional reasons* beyond them. Note that these are only the simplest counterfactuals, grounding IRCs between theories that share a common range of reasons, like our monist and pluralist views. The existence of counterfactuals seems less and less plausible in more complex cases, or for theories that are more distinct. Take the comparison between TU and a Kantian version of PAD, say. Suppose that TU is correct. And consider a fact like 'if a Kantian PAD were true, the reason not to lie to a friend would be equally as strong as our actual reason to bring about twenty-three happy cows'. The assumption that the world is populated by such more complex mind-independent counterfactuals quite definitely comes at considerable cost. And unless we have a positive story about what could ground them or why we should assume them, we're now basically just asserting what the sceptics deny.

Tarsney himself acknowledges this last difficulty. He defends his metaphysical account only for theories with 'common content'[17] or 'shared assumptions'.[18] So he concedes that 'comparability classes of normative theories may turn out to be few, small, and far between'.[19] But if this really is the most we can hope for in terms of mind-independent reason-counterfactuals, metaphysical accounts are at best rather weak. They only explain IRCs for a relatively small subset of theories. To remedy this shortcoming, proponents of such an account might combine their approach with other methods of comparisons or alternative theories of uncertainty.[20] They might hold that where metaphysical grounds are lacking, you ought to use a structural normalization principle to make comparisons, or that in such cases you simply ought to do what your favourite theory says. But these extensions seem *ad hoc*, or at least quite inelegant. Other things equal, we should prefer an account that delivers IRCs for a broader range of theories.

II.3. Absolutist accounts

So let me discuss a final approach. I'll refer to proposals of this kind as 'absolutist accounts'. On these accounts, moral theories make statements about the *absolute* strength of our reasons, and IRCs are grounded in these claims.

The most prominent version of this idea employs fitting attitudes. On this proposal, there's an attitude or set of attitudes such that the fact that you ought to choose *a* rather than *b* means that it's fitting to have these attitudes. For instance, it might mean that it's fitting to be disappointed if you chose *b*. Furthermore, these attitudes come in degrees, and the stronger your reason for doing *a* rather than *b*, the stronger the attitudes that are fitting. Finally, a complete moral theory must tell you not only what you ought to do, but also what *absolute* degrees of such attitudes are fitting. So it must tell you not

[17] Tarsney, 'Intertheoretic Value Comparison', p. 327.

[18] Tarsney, 'Intertheoretic Value Comparison', p. 332.

[19] Tarsney, 'Intertheoretic Value Comparison', p. 336.

[20] See Tarsney, 'Intertheoretic Value Comparison', p. 338.

https://doi.org/10.1017/S0953820819000165 Published online by Cambridge University Press

only that you ought to choose *a*, say, but also whether you ought to be slightly, or quite, or extremely disappointed, if you chose *b*. Consequently, there are infinitely many versions of any moral ordering. For example, there's Keyed Up TU, according to which you ought to maximize total welfare and be extremely disappointed if you made someone suffer a pinprick. And there's Calmed Down TU, according to which you ought to maximize total welfare but be only mildly disenchanted if you caused masses of people to be tortured. In this sense, theories make statements about the *absolute* strength of our reasons. This has been suggested by Ross, who said: 'The scale of a value function can matter … quite apart from issues raised by evaluative uncertainty. … Two linearly evaluative theories can disagree … concerning the degree of disappointment that is warranted.'[21] If all of this is true, IRCs can be grounded in theories' claims about attitudes. So for it to be the case that the moral reason you have for doing *a* rather than *b*, according to TU, is stronger than the moral reason you have for doing *b* rather than *a*, according to PAD, is for the attitude it would be fitting to have towards *a* and *b* if TU is true to be stronger than the attitude it would be fitting to have towards *b* and *a* if PAD is true. And whether or not this is so will depend on the versions of TU and PAD we consider.

What's nice about this proposal is that it doesn't presuppose any *extra* facts, like metaphysical reason-counterfactuals, beyond the facts implied by first-order theories. It's facts implied by the theories themselves that ground comparisons. But here too the virtue comes at a cost. The core absolutist assumption just seems dubious. There don't seem to be any facts about fitting absolute degrees of attitudes. If you're certain that TU is the correct moral ordering, say, it seems meaningless to wonder whether *all* your reasons could perhaps be (linearly) stronger than you thought, or whether it would be fitting for you to care (proportionally) more, or less, about *everything*. Suppose one person has Keyed Up TU-attitudes and another has Calmed Down TU-attitudes. They agree on all the kinds and *relative* strengths of attitudes. So whenever one of them is disappointed about someone's choosing *a* rather than *b* then so is the other, and whenever one of them is five times as disappointed about this than about someone's choosing *c* over *d* then so is the other. But all of the first person's attitudes are stronger in absolute terms. On the present proposal, at least one of them must be making a mistake, and misjudge the strength of their reasons. But this seems implausible. It seems that the first person is more emotional than the second, and that's that.

This isn't to say that people's attitudes aren't criticizable. Usually, if you get wildly furious at my being two minutes late or feel only a slight disenchantment about a genocide, your attitudes are unfitting. But this is only because you will usually have other attitudes that show you are getting the moral ranking of options wrong – considering a short delay as on a par with murder, or a genocide as on a par with a lie. If you had one of these attitudes, but had proportionally strong or weak attitudes about *everything* else, you'd not be misjudging anything. You'd be tragically emotional or pathologically indifferent. And since your life would be better if you cooled down or warmed up, you might have *state*-given reasons to work on your attitudes. But you wouldn't be getting any *fact* wrong. Or so at least it seems.

III. Ought-first

Let me take stock. If my suggestions in this brief overview were right, we have reasons to look for a more parsimonious account on which IRCs are not grounded in any extra

[21] Ross, 'Rejecting Ethical Deflationism', p. 765.

https://doi.org/10.1017/S0953820819000165 Published online by Cambridge University Press

mind-independent facts about how two theories compare (as on metaphysical accounts), or facts about the absolute strengths of our reasons (as on absolutist accounts), but are nonetheless sensitive to the content of our theories (unlike structural proposals). I think there is such an account. So let me turn to it now.

Let's start with the question about what it *is* for an IRC to hold. I've suggested that we don't have an intuitive grasp of what they mean. So we need to explicate what IRCs even say. But there's a final proposal in the literature that provides a more promising answer. It's an idea mentioned briefly by Ross.[22] Ross pointed out that we can understand IRCs in terms of facts about what you ought to do under uncertainty. On this proposal, for the moral reason you have for doing *a* rather than *b* according to TU to be stronger than the moral reason you have for doing *b* rather than *a* according to PAD, is just for it to be the case that if you find both theories equally likely (and these are the only theories you have credence in), you ought to do *a* rather than *b* in light of your uncertainty. I'll call this an 'ought-first' approach – where by the 'ought' I mean the subjective 'ought' that determines what you ought to do in light of your uncertainty. We arguably understand statements about what you ought to do in light of your uncertainty. So this provides an answer to the question about what *it is* for IRCs to hold. Moreover, as far as it goes, this answer is parsimonious in the sense that it doesn't in itself presuppose any mind-independent facts about how moral theories compare or about the absolute strengths of our reasons. So it's a promising start.

However, the ought-first approach also raises questions. For instance, there's a question about what conditions your moral theories and the ought-facts – the facts about what you ought to do in light of your uncertainty – must satisfy in order for them to imply unique, cardinal IRCs between all theories. And there's a question about whether or not, or to what extent, these conditions can be met. I've explored these questions in detail elsewhere.[23] So I won't pursue them here.

The ought-first approach also raises a more fundamental question. As it stands, the proposal might answer the question about what it *would be* for an IRC to hold. But the worries about object-level facts and their grounds still remain. They now simply arise for the ought-facts. A sceptic might agree that IRCs between TU and PAD could in principle be understood in terms of such ought-facts. But *are* there any facts about what you ought to do in light of your uncertainty between TU and PAD? And if so, what are they? If there are no ought-facts, then no IRCs hold. And if we don't know which ought-facts hold, we don't know which IRCs do. Or again, a sceptic might even concede that some ought-facts seem plausible. But what is it that grounds them? If we don't know what grounds the ought-facts, we don't know what ultimately grounds IRCs. So in order to have a more complete reply to scepticism, we must say more than the simple explication of IRCs in terms of 'ought' – even if this explication could technically work.

There are various options for an ought-first account, and some of them have already been considered. As indicated, Ross himself seems to endorse both an ought-first *and* an absolutist account. And indeed, perhaps we could somehow invoke fitting attitudes as grounds for ought-facts. But for the reasons I mentioned, I'm sceptical about this strategy.

[22] Ross, 'Rejecting Ethical Deflationism', p. 763.

[23] S. Riedener, 'Maximising Expected Value under Axiological Uncertainty: An Axiomatic Approach' (PhD dissertation, University of Oxford, 2015); Riedener, 'An Axiomatic Approach'.

https://doi.org/10.1017/S0953820819000165 Published online by Cambridge University Press

Another option would be to hold that the facts about what you ought to do under uncertainty are *fundamental* facts that aren't grounded in anything. So we could hold that it's a brute fundamental fact that you ought to bring about the new population rather than benefit the existing person. However, in saying this, we wouldn't have gained much ground over the sceptics. To the extent that the sceptics have doubted IRCs, they will be sceptical about fundamental ought-facts. And indeed, it does seem implausible that such facts should be brute. Perhaps there are some fundamental normative facts, such as that we're all morally equal. But facts about what you ought to do under uncertainty are highly complex. It seems hard to believe that *they* should be fundamental.[24]

Yet another option would be to go entirely subjectivist. We could say that there are no objective facts about what you ought to do when you're uncertain between TU and PAD. Rather, it all depends on you. *You* must have beliefs not only about the plausibility of moral orderings, but also about the possible relative strength of your reasons. You might have credence in the view that PAD is true and you have comparatively strong reasons to follow it – i.e. that you ought to give a lot of weight to PAD vis-à-vis TU under uncertainty. Or you might have credence in the view that PAD is true and you have comparatively weak reasons to follow it – i.e. that you ought to give little weight to PAD vis-à-vis TU. We might say you must believe in one or another version of PAD, at least *relative* to TU. And these beliefs will ground your ought-facts. For example, if you have credence in the view that PAD is true and you have comparatively strong reasons to follow it, then you ought to give PAD a lot of weight vis-à-vis TU in light of your uncertainty. So what you ought to do depends, radically, on what you believe you ought to do. This is a coherent version of an ought-first account.[25] But it has very radical implications. If there are no objective standards to distinguish reasonable ought-beliefs from unreasonable ones, we arguably can't speak of 'beliefs' in the first place. Belief presupposes a standard of correctness. So this entirely subjectivist proposal reduces IRCs to something like arational personal preference: the fact of an IRC holding between your theories would be a merely psychological fact about you. And this would not only imply that if you have no such preference, the theory of moral uncertainty cannot be action-guiding for you. It would also imply that you could permissibly assume that the reasons of pleasure would be 113.27 times stronger if beauty also had value, say. Indeed, it would imply that whenever you have *some* non-zero credence in a theory on which you ought to choose some option, then in light of your uncertainty it can be permissible for you to choose it. For instance, if you have *some non-zero* credence in the Nietzschean view that you're an *Ubermensch* permitted to do what you please, there are no grounds for criticizing your judgement that in light of your credences you may do what you please. But these are surely unfortunate results. Some ought-statements simply seem false, and some ought-beliefs unreasonable. IRCs aren't entirely subjective.

IV. Constructivism

Let me try to do better. *Why* is it unreasonable to believe – or make ought-judgements to the effect – that the reasons of pleasure are 113.27 times stronger on the pluralist

[24] For a related worry, see Tarsney, 'Intertheoretic Value Comparison', p. 327.

[25] If I understand her correctly, this is roughly the line taken by A. Hicks, 'Moral Uncertainty and Value Comparison', *Oxford Studies in Metaethics*, vol. 13, ed. R. Shafer-Landau (Oxford, forthcoming).

https://doi.org/10.1017/S0953820819000165 Published online by Cambridge University Press

theory? In particular, how can this be if there's no independent metaphysical fact that makes this comparison false? The key idea, I suggest, is that we can ground IRCs in *epistemic norms*. There are epistemic norms that are plausible independently of the problem of intertheoretic comparisons, but constrain the IRCs you can make. We can understand them as holding prior to any IRCs, and as grounding IRCs in a constructivist manner.

To illustrate what I mean, let me first give some examples of the kind of norm I have in mind, and of how they can be constraining. One type of norm might be synchronic norms concerning your credence distribution at any time. A good candidate of this kind is

> *Simplicity: ceteris paribus*, you should favour simpler credence distributions over more complex ones.

It's difficult to spell out precisely what 'simplicity' means, but we arguably have an intuitive understanding of it. So suppose that Simplicity holds. Then it can constrain the IRCs you can reasonably make. The credence distribution on which the reasons of pleasure are equally strong on the monist and the pluralist theory is arguably simpler than that on which their ratio is 113.27, or anything other than 1. So if Simplicity is true, and if you have no reason to believe anything else, you should favour this simple IRC. More generally, you should *ceteris paribus* believe that the reasons shared by overlapping theories are equally strong on both. You should make ought-judgements that imply IRCs of this form.

Other candidate norms are diachronic ones concerning the evolution of your credences over time. Consider epistemic conservatism, the idea that you should not change your beliefs in the absence of any reason to do so.[26] An implication of this idea for how to deal with new evidence might be put as

> *Conservatism*: if you encounter new evidence, then of the possible changes to your credences that accommodate this evidence you should *ceteris paribus* favour less radical over more radical ones.

This norm too constrains your IRCs. Suppose you've so far believed in the pleasure-theory, but now encounter some evidence for the value of beauty. The least radical way to accommodate this evidence is to adopt some positive credence in reasons of beauty, but to leave your beliefs about pleasure unchanged. Any IRC on which the reasons of pleasure are stronger, or weaker, on the pluralist theory than on the monist one would suggest that you may so far have misjudged their strength. So it would imply that you'd have to change your mind about pleasure-reasons if you came to accept reasons of beauty besides them. But epistemic conservatism says there's a presumption in favour of not changing your mind, or believing you were wrong, without any positive grounds. So if Conservatism is correct, and if you have no positive countervailing reason, you should believe that the pleasure-reasons are equally strong on both theories. More generally, you should *ceteris paribus* not change your beliefs about some given reasons in the face of evidence for *additional* reasons besides them.

[26] See e.g. R. Chisholm, 'A Version of Foundationalism', *Midwest Studies in Philosophy* 5 (1980), pp. 543–64; J. Kvanvig, 'Conservatism and its Virtues', *Synthese* 79 (1989), pp. 143–63; K. McCain, 'The Virtues of Epistemic Conservatism', *Synthese* 164 (2008), pp. 185–200.

https://doi.org/10.1017/S0953820819000165 Published online by Cambridge University Press

As a third candidate norm, consider

> *Coherence: ceteris paribus*, you should favour more coherent credence distributions over less coherent ones,

where coherence is understood roughly as the degree to which your beliefs are mutually supportive. Such a norm can also constrain your IRCs. It can do so for instance if you have a positive error theory about why you might have been mistaken about morality. Suppose again you've so far believed in the pleasure-theory, but now encounter some evidence for the value of beauty. And suppose you have a belief, conditional on the pluralist theory, about why you've long missed its truth. For example, you believe that if beauty also had value, you would simply have been insensitive to its particular worth. This explanation suggests that even if the pluralist theory is true, you've never made any mistake with respect to your reasons of pleasure. So it arguably coheres best with the IRCs on which the reasons of pleasure are the same on both theories. Any alternative credence distribution would suggest that you haven't just overlooked your reasons of beauty, but also misjudged your reasons of pleasure. And this wouldn't square well with your own simple error theory.

So very roughly, these norms suggest that without any explanation, you shouldn't assume that you've always systematically and radically misjudged the strength of your everyday paradigm reasons. And they imply that you should more readily assume you may have misjudged some reasons if you have an explanation for why and how you may have done so, or if these reasons are less mundane and pervasive. This seems intuitively plausible. But Simplicity, Conservatism and Coherence might be false, or not quite correct as I've stated them, or there might be other and more important norms besides them.[27] My aim is not to argue for these precise norms. I'm happy if it's plausible that *some* such epistemic norms hold, and that they can constrain the IRCs or ought-judgements you can reasonably make. If that's so, we can invoke a form of constructivism to ground IRCs. We can understand truth about IRCs as the outcome of ideally reasonable deliberation – in terms of principles like the above – about what you ought to do in light of your uncertainty. By comparison, consider the view that truth in first-order moral theory is simply the result of an ideal process of systematizing our pre-theoretical moral beliefs.[28] On this view, it's not that there's some independent Platonic realm of moral facts, and that norms like simplicity and coherence are best at guiding us towards it. Rather, the principles are first, and 'truth' is simply the outcome of the principles. We can invoke a similar kind of constructivism about IRCs. On this view, principles like Simplicity, Conservatism and Coherence are not justified in virtue of their guiding us towards an independent realm of ought-facts or IRCs. Rather, they help constitute this realm.

[27] For objections to epistemic conservatism, see e.g. D. Christensen, 'Conservatism in Epistemology', *Noûs* 28 (1994), pp. 69–89; H. Vahid, 'Varieties of Epistemic Conservatism', *Synthese* 141 (2004), pp. 97–122. Perhaps there's a quasi-intuitionist rationale for Conservatism, or something like it, with respect to moral beliefs: perhaps we have a (fallible, but non-negligible) faculty to detect first-order moral facts, and this gives us reasons to treat our moral beliefs as evidence, or to revise them as little as possible when we need to.

[28] See e.g. J. Rawls, 'Kantian Constructivism in Moral Theory', *Journal of Philosophy* 77 (1980), pp. 515–72; C. Korsgaard, *The Sources of Normativity* (Cambridge, 1996).

https://doi.org/10.1017/S0953820819000165 Published online by Cambridge University Press

So this provides an answer to why some ought-facts or IRCs hold. It's *not* because of mind-independent metaphysical facts about how theories compare, or how strong certain reasons would be if we had them. It's simply because of facts about how to respond reasonably to moral evidence or have reasonable moral beliefs. Ultimately, we might say, it's because of facts about *us* – about why we might have been wrong about morality, and by how much and in what way, and so on.

V. Applications and clarifications

Admittedly, the case of our two theories was exceptionally simple. So let's see how these epistemic principles could be applied in more complex cases. Consider our initial case again. Suppose you've long believed in total utilitarianism, but have recently come to doubt it, and now slightly prefer some form of person-affecting deontology. What IRCs (or ought-judgements) would it be most reasonable for you to make? This depends on the version of PAD you have credence in and on your other beliefs. Suppose you believe in a welfarist PAD on which morality is fundamentally about doing good for others, but the relevant 'others' are only those who already exist. Then there's a real overlap between TU and your PAD. According to TU, you have all the reasons that you have according to PAD – reasons to benefit existing others – but also some additional reasons beyond them. So on this interpretation, the least radical change in your credences and the most simple ultimate credence distribution will be such that your reasons to benefit existing people are the same on both theories. Unless you have some additional beliefs that could render other beliefs more coherent, this IRC will be most reasonable in light of the above principles.

However, suppose you believe in a Kantian PAD on which morality is fundamentally about interpersonal respect and concern for the autonomy of agents, and only marginally about beneficence. On this view, you do have reasons to do good for others, if these others exist and want to be benefited by you. But these reasons are relatively weak, and easily outweighed by your reasons of self-interest, or by your negative reasons not to actively harm or lie or break promises. The comparison between TU and this theory is more complex. Simplicity might favour a credence distribution on which the reasons of beneficence accepted by both theories – towards existing others that want to be helped – are the same on both views. But these reasons are very weak within PAD. So this comparison will imply that the relevant *other* reasons of PAD are massively stronger than our standard reasons of TU. It thus constitutes a radical departure from your original beliefs. You've so far believed that you have no such extra PAD-reasons. But on this comparison, you might have *extremely strong* reasons of this kind. So the comparison suggests that you might have misjudged *massively* all your reasons of self-interest, and your reasons not to harm or lie or break promises. Conservatism might thus favour a credence distribution on which these reasons of beneficence are stronger on TU than on PAD. Such a comparison suggests that you might have misjudged your reasons of beneficence, as well as your other PAD-reasons. But it doesn't suggest that you might have been so horrendously wrong about these other PAD-reasons. So the overall extent to which you represent yourself as possibly having misjudged your reasons might be smaller on this second comparison than on the first. And these implications of Conservatism might have to be balanced against those of Simplicity.

But suppose you also have some pertinent non-normative beliefs. Consider the fact that most people intuitively feel we have very weak reasons to create new people,

https://doi.org/10.1017/S0953820819000165 Published online by Cambridge University Press

compared (say) to our reasons not to actively harm an existing person. If TU is true, there must arguably be some explanation for why we're getting this wrong. Suppose you have a belief about that. Suppose you believe that conditional on TU, we're very bad at imagining and appreciating all the good we could do for merely possible people. This suggests that we might grossly underestimate our reasons to create these people, relative to our other reasons. So in this case, Coherence might favour IRCs on which you know the strength of your reasons not to harm existing people – i.e. on which TU and PAD agree on *these* – and on which you have correspondingly strong reasons to create new people if TU is true. But you might have some other explanation. Suppose you believe it was evolutionarily advantageous for us to develop a strong reluctance to actively harm, over and above our inclination to benefit or our capacity for empathy. You believe that conditional on TU, this reluctance doesn't track any moral truth, and that we thus mistake the noisy firing of these emotional neurons for additional moral reasons. This suggests that we might overestimate our reasons not to harm existing people, relative to our reasons to create new ones. So in this case, Coherence might favour IRCs on which you roughly know the strength of your standard reasons to do good – i.e. on which TU and PAD roughly agree on *these*, where they both accept them – and on which you have correspondingly stronger reasons not to harm existing people if PAD is true. And whatever explanation you prefer, the implications of Coherence might again have to be balanced against those of Simplicity, say.

All of this is only a rough sketch of how norms like Simplicity, Conservatism and Coherence could operate in more complex cases. No doubt, what precisely these principles imply will be a complex question. Indeed, there might often not be a precise comparison that comes out as most reasonable. It might square best with these principles to assume that some theories are only *roughly* comparable. And perhaps there are even cases where the three principles I've considered fail to be constraining at all – for instance, when you're uncertain between TU and the theory on which the only valuable thing is your bicycle tyre. I've not suggested that I have a ready method for determining all comparisons. Given that the study of IRCs is in its infancy, it would be surprising if we had that. But I hope the discussion indicates that applying these norms can be fruitful. In particular, they seem to give us more resources to make IRCs than metaphysical accounts allow. As I've shown, norms like Conservatism and Coherence have non-trivial implications for the comparison between TU and PAD, in light of our prior beliefs or in light of our beliefs about our beliefs in the form of an error theory. These resources don't seem to be available for metaphysical accounts. Indeed, as I've suggested, it's unclear what a mind-independent basis for IRCs between TU and PAD could be. So there's reason to be more optimistic that IRCs hold not only for 'few and far between' theories.

One final point is worth mentioning. Note that in light of *these* principles the relevant ought-facts might be different from person to person. Most notably, it depends on your priors which IRCs effect the least radical changes in your credences. So it seems that if the above principles hold, there cannot be a *universal* truth of the form 'the moral reason you have for doing *a* rather than *b*, according to TU, is stronger than the moral reason you have for doing *b* rather than *a*, according to PAD'. Whether or not this statement is true seems different from person to person.

We might be worried about this. And perhaps we could avoid it. Perhaps we can define a sort of 'ideal deliberation' or find a set of principles which rules out these interpersonal differences. However, I see no reason to be worried. We should accept that IRCs are only true relative to some set of prior beliefs. Recall that on the ought-first

https://doi.org/10.1017/S0953820819000165 Published online by Cambridge University Press

approach, IRCs only mean that you ought to choose some options rather than others. And it's not surprising that what you ought to do – in the less than fully objectivist sense we're considering – should depend on your prior beliefs. This doesn't mean that one and the same IRC-proposition is true for one person but false for another. This would be dubious. Rather, if different ought-facts hold for different people who are uncertain between TU and PAD, we can understand this as meaning that it's reasonable for these people to believe in different versions of these theories. Talk of 'different versions' of TU, say, doesn't presuppose absolutism. We can assume (*pace* Ross[29]) that the versions differ in nothing 'apart from issues raised by evaluative uncertainty'. On this assumption, the only difference between two versions of TU concerns what you ought to do in light of uncertainty about them. If this is so, their difference will only be apparent *relative* to some fixed version of PAD, say. And two people who have the same credences in the same orderings and for whom the same ought-facts hold cannot be further distinguished – as they can on absolutism, where all of one person's theories might be keyed-up versions of the theories of the other. But all of this seems plausible. Different ought-facts can be true relative to different people, since it can be reasonable for them to believe in what are – in this thin sense – different versions of their theories.

VI. Conclusion

I've sketched the foundations of an account about what IRCs mean, about what grounds them and which of them are true. On this account, we understand IRCs in terms of facts about what you ought to do under uncertainty. And we explain in a constructivist manner why some such facts hold, and which of them do. If I'm right, this proposal has several virtues. It can explain why the ought-facts are neither brute nor entirely subjective. It does so in a way that's metaphysically more parsimonious than other accounts. And yet it delivers plausible and strong results about which IRCs hold: it can capture their content-sensitivity, and gives us resources to account for IRCs even between very distinct theories.

Many questions remain open. What are the most plausible epistemic norms? How far will they get us in grounding IRCs in the manner I've sketched? And can we (or need we) somehow ground these norms in turn? These questions are beyond the scope of this article. But I hope I've shown that constructivism about intertheoretic comparisons is an idea worth exploring.[30]

Author ORCIDs. Stefan Riedener, 0000-0002-7315-1016.

[29] Ross, 'Rejecting Ethical Deflationism', p. 765.

[30] For enormously helpful conversations and feedback, I thank John Broome, Hilary Greaves, Adam Lovett, William MacAskill, Christian Tarsney, Teruji Thomas, and audiences at the Global Priorities Institute Seminar and the ISUS 2018 Karlsruhe. I also thank the Global Priorities Institute for hosting me in Oxford while I was working on this article.

Cite this article: Riedener S (2019). Constructivism about Intertheoretic Comparisons. *Utilitas* **31**, 277–290. https://doi.org/10.1017/S0953820819000165