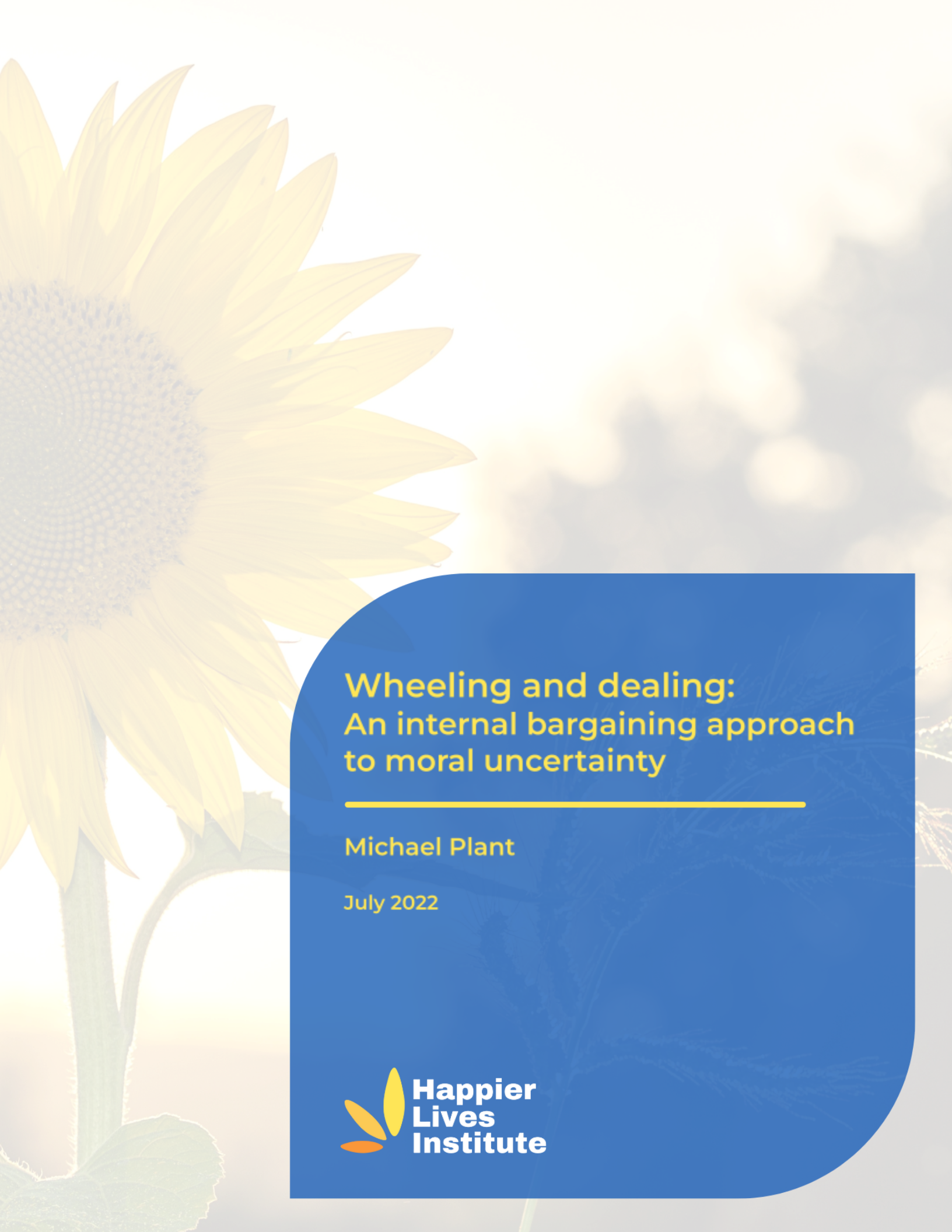

# Wheeling and dealing:
## An internal bargaining approach to moral uncertainty

Michael Plant

July 2022

Happier Lives Institute

# Contents

# Summary

In this post, I explore and evaluate an approach to moral uncertainty that draws on a metaphor of *internal bargaining* (IB). On this approach, the appropriate choice under moral uncertainty is the one that would be reached as the result of bargaining between agents representing the interests of each moral theory, who are awarded your resources in proportion to your credence in that theory. This has only been discussed so far by Greaves and Cotton-Barratt (2019), who give a technical account of the approach and tentatively conclude that the view is inferior to the leading alternative approach, *maximise expected choiceworthiness* (MEC). I provide a more intuitive sketch of how the internal bargaining works, and do so in a wide range of cases. On the basis of the cases, as well as considering some challenges for the view and its theoretical features, I tentatively conclude it is superior to MEC. I close by noting one implication relevant for effective altruists: while IB would provide a justification for something like *worldview diversification*, MEC instead pushes towards a (fanatical) adherence to *longtermism*.

***Notes to reader:*** *(1) I'm deliberately writing this in a fairly rough-and-ready way rather than as a piece of polished philosophy. If I had to write it as the latter, I don't think it would get written for perhaps another year or two. I'll shortly begin working on this topic with Harry Lloyd, an HLI Summer Research Fellow, and I wanted to organise and share my thoughts before doing that. (2) this can be considered a 'red-team' of current EA thinking.*

# 1. Introduction

When philosophers introduce the idea of *moral uncertainty* - uncertainty about what we ought, morally, to do - they often quickly point out that we are used to making decisions in the face of *empirical uncertainty* all the time.[1]

Here's the standard case of *empirical uncertainty*: it might rain tomorrow, but it might not. Should you pack an umbrella? The standard response to this is to apply *expected utility theory*: you need to think about the chance of it raining, the cost of carrying an umbrella, and the cost of getting wet if you don't carry an umbrella. Or, more formally, you need to assign credences (strengths of belief) and utilities (numbers representing value) to the various outcomes.

[1] For an article-length overview of moral uncertainy, see Bkyvist (2017). For a book length discussion of the state-of-the-are, See MacAskill, Ord, Bkyvist (2020)

Hence, when it's pointed out that we're *also* often uncertain about what we ought to do - should we, for example, be consequentialists or deontologists? - the standard thought is that our account of *moral* uncertainty should probably work much like our account of *empirical* uncertainty. The analogous account for moral uncertainty is called *maximise expected choiceworthiness* (MEC) (MacAskill, Ord, Bkyvist, 2020). The basic idea is that we need to assign credences to the various theories as well as a numerical value on how *choiceworthy* the relevant options are on those theories. The standard case to illuminate this is:

> ***Meat or Salad*:** You are choosing whether to eat meat from a factory farm or have a salad instead. You have a 40% credence in View A, a deontological theory on which eating meat is seriously morally wrong and a 60% credence in View B, a consequentialist theory on which both choices are permissible. You'd prefer to eat meat.

Intuitively, you ought to eat the salad. Why? Even though you have less credence in A than B, when we consider the relative stakes for each view, we notice that View A *cares* much more about avoiding the meat. Hence, go for the salad as that maximises choiceworthiness.

MEC is subject to various objections (see MacAskill, Ord, Bkyvist, 2020 for discussions and ultimately a defence of the view). First, the problem of intertheoretic comparisons: there is no broadly accepted way to make comparisons of choiceworthiness across theories. For instance, a deontological theory might say murder is morally worse than theft, but say nothing about how much worse the former is; how would we compare this against a utilitarian theory that compares outcomes in units of welfare losses and gains?

Second, MEC also leads to problematic forms of *fanaticism*, where a theory in which one has very low credence, but according to which there is an enormous amount at stake, can dictate what one should do. For instance, consider:

> ***Lives or Souls*:** You're about to donate to the Against Malaria Foundation and you know that your donation would save one child's life in expectation. Someone then shows you evidence that the Against Hell Foundation reliably converts one person to a certain religion and thereby saves their soul from eternal damnation. You have almost no credence in that religion, but know that saving a soul is, on that religion, incomparably more valuable than saving one life. You realise that MEC implies you ought to give to the Against Hell Foundation instead, but this seems wrong.

Relatedly, MEC is worryingly fanatical not only about *what* you should do, but *how much* you should do (MacAskill, Ord, Bkyvist, 2020, chapter 2). Consider:

> ***Partial Singerian*:** you have a 10% credence in the view advocated by Peter Singer that citizens of rich countries ought to give a large proportion of their resources to help those in poverty. The remainder of your credence is in common-sense views of morality on which doing this, whilst laudable, is not required.

What does MEC conclude? Well, the Singerian view holds there is a lot at stake if you give - think of all the lives you could save! - so it seems that you are pressed to accept a very demanding moral theory anyway. What seems odd and objectionable about MEC is the way that, in a certain sense, one small part of you can 'bully' the rest of you into doing what it wants if it cares enough.

Are there any alternatives to MEC? The other commonly-discussed approach is *My Favourite Theory* (MFT), which tells you to follow the moral theory you think is most likely to be true. But MFT seems worse than MEC.

First, MFT is vulnerable to *individuation*, how theories are divided up (Gustafsson and Torpman, 2014, section 5; MacAskill and Ord, 2018, pp.8-9). Note that for *Meat or Salad*, MFT currently recommends you choose meat as you have 60% in that theory. However, you then realise there are two version of consequentialism (e.g. act-consequentalism and rule-consequentialism) and you have a 30% credence in each. Now MFT tells you to eat the salad. So, to function at all, MFT needs a theory of option-individuation that is robust to these sort of objections.

Second, even supposing such a theory could be found, MFT is also insensitive to the stakes. If you had a 51% credence in a theory on which X is slightly better than Y, and a 49% credence in another theory on which X was enormously worse than Y, it would still recommend X.

What's more, neither MEC nor MFT can account for the strong intuition many people have that, in the face of our moral uncertainty, we should sometimes 'split the pot' and diversify our resources across different options. Prominently, Holden Karnofsky of Open Philanthropy has advocated for *worldview diversification*, which involves "putting significant resources behind each worldview [one] considers plausible". As a motivating case, consider:

> ***Poverty or Robots*:** You have 60% of your credence in a *total utilitarian* moral theory on which the greatest priority is preventing existential risks to humanity, such as those posed by unsafe artificial intelligence. But you have 40% of your credence in the *person-affecting view* on which morality is about "making people happy, not making happy people" (Narveson, 1967) and that, according to this view, helping those in global poverty is the priority. On the *total utilitarian* theory, using money to reduce poverty is good, but hundreds of times less cost-effective than funding existential risk.

MEC and MFT will both recommend that all the resources go towards existential risk reduction. In fact, MEC would still recommend this *even if* you only had 1% credence in the Total Utilitarian view (cf. Greaves and Ord, 2019). But many people feel the right thing to do is, nevertheless, to split your resources rather than give the whole pot to either cause.[2]

That we don't have a justification for 'worldview diversification' is a problem. Many people practice worldview diversification and seem to believe it is an appropriate response to moral uncertainty. At the very least, we want an account of moral uncertainty that could make sense of this.

Those are some challenges for both MEC and MFT. I will now suggest an approach to moral uncertainty that might do better. This has been introduced by Greaves and Cotton-Barratt (2019) who call it *bargaining-theoretic*, but I prefer, and will use, the more intuitive *internal bargaining* (IB). As noted above, on this approach, decision-making under moral uncertainty is modelled as if it were a case of bargaining between different parts of the agent, where each sub-agent is committed to a moral view and resources are awarded to sub-agents in proportion to the credence the agent has in those views. In other words, we can say the appropriate response to ethical uncertainty is what is decided in the moral marketplace.

---

[2] Now, there is a justification for diversifying that does not appeal to moral uncertainty. Suppose you think that the interventions exhibit diminishing marginal returns, so you do the most good by funding one intervention and then switching to another. The total utilitarian might believe that spending a little bit of money on AI safety research goes a long way, but once all the low-hanging fruit has been picked it's more impactful to give to poverty. Of course, the facts might be different. You could think that spending extra money on existential risk reduction is always going to do more good than spending money to reduce poverty. In this case, the total utilitarian would tell you not to diversify.

But notice that this discussion of whether to diversify makes no reference to moral uncertainty and takes place within a single moral theory: we might call this *intra*-worldview diversification. What we don't yet have is a justification for splitting our resources based on uncertainty about morality, what we can call *inter*-worldview diversification. I take it when most people appeal to worldview diversification they have the 'inter' version in mind, and this is how I will use the term here.

Greaves and Cotton-Barrett (2019) provide a technical treatment of how this view would work which draws of bargaining theory in economics. It is sufficiently technical that I expect only mathematically sophisticated readers will come away from it feeling confident they understood how internal bargaining would work and are able to form a view on how plausible an approach it is. Here, I aim to provide a commonsense account of *roughly* how the view works by relying on nothing more than our ordinary intuitions about bargaining. Because we are so used to bargaining in ordinary life, this should carry us a long way. Hence, I don't aim to say *exactly* how the view works - that requires the sort of further detail in Greaves and Cotton-Barrett (2019). I'm not sure I have a substantially different view of how internal bargaining would play out - although it's possible I'm misread their account. However, I *am* more enthusiastic about how suitable internal bargaining is as a response to moral uncertainty.

Here's the structure and summary of what follows.

Section 2 offers an taxonomy of different scenarios and provide an intuitive sketch of how internal bargaining would play out in each; I conclude the approach gives plausible results.

Section 3 raises various problems and open questions for the view.

Section 4 reviews the theoretical features of the view and argue it seems to capture most of what we want from an account of moral uncertainty.

Section 5 note a practical implication of which approach to moral uncertainty is chosen for effective altruists: IB leads to *worldview diversfication*, whereas MEC pushes us towards *longtermism,* the view that the priority is improving the long-term future.

# 2. What would happen in the moral marketplace?

Here's a familiar scenario. You and a friend are deciding where to go for dinner. You prefer Italian food. They prefer Indian food. But you both prefer to spend time with each other. What do you do? Maybe you compromise on a third option, Mexican. Maybe you take it in turns so each of you can go to your preferred option. But maybe there's no compromise that works for you both and so you each do your own thing. This is deeply familiar and very intuitive and, to give it a fancy term, it is *inter*personal bargaining. We engage in interpersonal bargaining in our personal lives, our work lives, in politics, and so on.

Suppose we conceive as moral uncertainty as the result of *intra*personal bargaining between different theories, which I'll refer to as *internal bargaining*. You divide yourself up into sub-agents, each of which is fully committed to a moral theory. You then allocate your resources - your money and time - to these sub-agents in proportion to how strongly you believe in each and do what your sub-agents collectively decide.

To see how this might work I'm proposing a taxonomy the various scenarios and what would happen in each.[3] I'll only consider that the agent has credence in two moral theories.

I'll go through these one at a time, starting with with three scenarios where you can divide the resources between your sub-agents. What will differ is whether the sub-agents have convergent priorities (you and I want the same thing), conflicting priorities (we want different things) or unrelated priorities (it doesn't make any difference to me if you get more of what you want, and *vice versa*). I'll then consider what happens if resources aren't divisible.

***Table 1:*** *Taxonomy of moral uncertainty scenarios*

| | **Divisible resources** | **Indivisible resources** |
|---|---|---|
| **Convergent priorities** | Unity | Unity |
| **Conflicting priorities** | Intrapersonal moral trade | Intrapersonal moral trade (over time)<br><br>Or, perhaps, overpowering |
| **Unrelated priorities** | Split the pot<br>(aka worldview diversification) | N/A |

**Note:** There is no 'indivisible resources, unrelated priorities' scenario. If resources aren't divisible, then inevitably one agent getting their preferred option is at the cost of the other achieving their preferred option.

[3] I hope this is the full range of options, but I might have missed something.

## Divisible resources, convergent priorities

**Example**: *Charitable Consensus*

Both moral views agree on which charity would do the most good.

**Result**: *Unity*

All the resources go to one charity. There are no disagreements or any need for bargaining.

## Divisible resources, conflicting priorities

**Example:** *More or Fewer*

On theory A, life-saving interventions are the priority. On theory B, the Earth is overpopulated and the priority is planning interventions to reduce the population size. However, both A and B each think that funding a third, life-saving interventions, e.g. alleviating poverty, is nearly as effective as their own top choice. You have equal credences in theories A and B.

**Result:** *Intrapersonal moral trade*

The sub-agents realise that, if they each pursue their own preferred option, they will effectively cancel each other out - A would increase the total population and B would reduce it. Hence, by the lights of their own theory, they would each prefer it if they collectively funded poverty reduction, so that's what they choose to do.

(In fact, I'm oversimplifying here. It won't always be the case that agents agree to engage in moral trade. That will depend on their relative resources and how good they think the available options are. For instance, if theory A has £1000 and theory B £1, then theory A might prefer putting all its money towards saving lives, and having B slightly counteract that money, than agreeing to a compromise. In this case, the agents will *split the spot*, as I elaborate on in the next example.)

## Divisible resources, unrelated priorities

**Example:** *Poverty or Robots* (as given above)

You have 60% of your credence in a *total utilitarian* moral theory on which the greatest priority is preventing existential risks to humanity, such as those posed by unsafe artificial intelligence. But you have 40% of your credence in the *person-affecting view* on which morality is about "making people happy, not making happy people" (Narveson, 1967) and that, according to this view, helping those in global poverty is the priority. On the *total utilitarian* theory, using money to reduce poverty is good, but hundreds of times less cost-effective than funding existential risk.

In this case, we might suppose, the sub-agents have effectively unrelated priorities: money to existential risk doesn't really impact poverty, and vice versa. What's more, the sub-agents can't find any scope for moral trade with the other: they each conclude that the best option, by their own lights, would be to do their own thing.

**Result:** *Split the pot aka worldview diversification*

Each sub-agent allocates all of its resources to its preferred option. In other words, the outcome is effectively *worldview diversification*, so we've identified a straightforward way of justifying this.

There are a couple of other observations to make here. Because IB leads to pot-splitting when theories disagree about the priority, it also offers a very natural way to resist *fanaticism*. Recall *Lives or Souls* earlier, where you had a tiny credence in a view on which saving souls was the priority. Yet, because you only have a tiny credence on that view, that sub-agent only gets allocated a trivial amount of resources. Hence, IB is *non-fanatical*: it doesn't ignore fanatical views altogether, but these have little impact in decision-making precisely because resources are awarded in proportion to credences and the agent has so little credence in them. This strikes me as a serious advantage of IB over MEC. If we take MEC seriously, we'd seemingly need to account for all sorts of 'weird' fanatical views. IB safely contains them.[4] I'll return to whether this *lack* of fanaticism may be a problem in a later section.

For similar reasons, IB also seems able to defuse the issues that MEC faces about demandingness. Recall the *Partial Singerian* case from before, where you have a 10% credence in the view you should give away your resources so long as they will make others better off, and a 90% credence in common-sense morality. There's no obvious bargain to be struck here, so we might imagine they

[4] It also safely handles the challenge of infectious comparability (MacAskill, 2013)

would just allocate their own share of the pot as they see fit. Roughly then, the Singerian sub-agent would give away all of their 10% share and the common-sense agent a little bit of theirs, with the result that the person ends up giving around (perhaps a bit over) 10% to charity.

Similarly, this seems a good response to the demandingness of morality, and far more palatable than the response given by MEC or MFT. The former pushes us to give all our spare resources away, even if we have little credence in the Singerian theory. The latter cannot account for the non-trivial Singerian belief that we ought to do as much as we can to help others. What's more, IB is sensitive - perhaps even respectful - of our credence: those who are very sympathetic to the Singerian view that morality is demanding will still conclude they should do lots, but those are less sympathetic are not pushed to do so.

Let's now turn to cases where you don't seem to be able to divide resources between the sub-agents. Again, cases can vary by how aligned the sub-agents' priorities are.

## Indivisible resources, convergent priorities

**Example:** *Cake or Death*

A runaway trolley is on course to run over someone tied to a track. You can pull a lever to switch the trolley to another line, but if you do, it will squash a nice piece of chocolate cake you were hoping to eat. Both moral theories agree you ought to pull the lever.

**Result:** *Unity - you pull the lever*

The salient difference, in contrast to the cases above, is that you can't split your resources; you have to choose between the options. However, as with *Charitable Consensus*, this is straightforward because both parties agree.

## Indivisible resources (at a time), conflicting priorities

**Example:** *Meat or Salad*

You are choosing whether to eat meat from a factory farm or have a salad instead. You have a 40% credence in View A on which eating the meat is seriously morally wrong and a 60% credence in View B on which each choice is permissible. You'd prefer to eat meat.

This is a more interesting case for internal bargaining and I'll consider four ways this could go.

## 1. Try to split the pot

An initial thought, taking inspiration from *Poverty or Robots*, is that you should try to split the pot: you could ask the restaurant to give you 40% of a normal meat order and 60% of a normal salad order. But, leaving aside how impractical this is, it seems to get the wrong answer: on View A, this is still much worse than if you'd only had salad, but only barely better on View B. This is a poor compromise because it is insufficiently sensitive to the stakes.

## 2. Lottery

Another possible option would be to use a lottery, e.g. there's a 40% chance that View A wins and you eat the salad and a 60% chance that View B wins and you eat the meat. This is similarly unappealing, again because it doesn't account for the stakes.

## 3. Intrapersonal bargaining overtime

However, a more creative option, along the lines of *More or Fewer?* is open: the sub-agents can strike a bargain, but in this case, they do so *over time* (aka *intra*personal *inter*temporal moral trade). To see this, let's first consider how an interpersonal version of this case could play out:

> ***Vegan friend*:** Two friends would like to meet for dinner. A has a really strong preference for vegan food, and will only eat at vegan restaurants. B would prefer to eat meat, but really isn't that fussed either way.

Plausibly, the real-life bargain is that they would agree to keep meeting, but only do so at vegan restaurants. And that could be the end of the story. However, we can also imagine that B would negotiate for something in return: "Hey, we always go to the restaurants you want - and that's fine, you know I don't care what I eat. But next time we go out for drinks, I'm choosing the bar". Exactly whether and what B negotiates for in return, and what they end up agreeing to, will depend on what A and B each care about.

Hence, in *Meat or Salad*, we can imagine, metaphorically speaking, that View A would protest to B about how it really matters to them that you end up ordering the salad. Plausibly, B would agree that they order the salad, but negotiate for something in return *at a later date*.

What might B negotiate for? Admittedly, this is much less intuitive to think about in cases of moral theories than people, but it's not impossible to sketch a result that they would agree to. Suppose theory A is a deontological theory on which there is a very strong duty not to harm others, but a relatively weak duty to benefit others (i.e. of beneficence). Theory B is classical utilitarianism. Potentially, the bargain would be that you don't eat meat but, in return, you end up doing more charity work than you otherwise would have.

A key point to note is that internal bargaining - like MEC - is able to get the intuitively right result about what to do in *Meat or Salad*. It does this by allowing each sub-agent to be sensitive to the stakes *according to their own theory* and then bargain to get more of what they want. It tells a different story of how moral uncertainty works and requires us to be a bit inventive in thinking about how internal bargaining might play out.

## 4. Overpower

However, there's (at least) one more option. If we're thinking about ordinary interpersonal bargaining, a live option is for the stronger party to force the weaker to do what it wants. To extend the metaphor, we might suppose the 'stronger' theory, i.e. the one if which you have more credence, could just 'overpower' the weaker one and pick its preferred option. In this case, IB would function much like My Favourite Theory (MFT) and theory B would 'win' and you would eat the meat.[5]

The raises the question of whether IB would function like MFT *all of the time.* If so, that would be a poor result for IB, given already noted problems with MFT. It seems we should say something like this: we suppose your sub-agents are entitled to their 'share' of your current *and* future resources - your resources being your money and time. Hence, while theory B could perhaps impose their will in *this* situation, the foreseeable result is that, in turn, A won't cooperate with B in some later situation where B would really want A's resources. Therefore, B may well conclude it's in their own best interest to 'play nice': whilst they'd rather eat the meat and could insist on it, because they really don't mind, they strike a bargain with A and ask for something in return. As such, it seems overpowering would be more of the exception than the rule.[6]

[5] IB behaves like MFT because there are only two theories at play. If we accounted for the fact that our credence could be split amongst many theories, then the metaphorical result would be the option with the largest proportion of sub-agents in favour would be the one that got chosen. In this case, IB would work like an alternative approach to moral uncertainty in the literature called *My Favourite Option* that I have not raised so far. The obvious objection to this approach is that it is insensitive to the stakes.

[6] This mirrors the result that in a *one-shot* Prisoner's Dilemma, it is rational for each party to defect, but that in a *repeated* Prisoner's Dilemma, the rational strategy is to cooperate.

I don't think it would be useful to discuss at much greater length how the bargaining could play out: it is only metaphorical and there is scope to 'tweak' the metaphor and get different outcomes - of course, we would want to try to provide some rationale for these tweaks.[7]

# 3. Problems and open questions

IB seems to have handled the cases well so far - there wasn't a scenario where it clearly got the intuitively 'wrong answer'. Let's turn to problems and open questions for the view. I'll only discuss these briefly; it's beyond the scope of this essay to solve them.

## Can IB account for the value of moral information?

In cases of empirical uncertainty, the value of information is the amount a decision-maker would pay to get closer to the truth. For instance, an ice-cream seller might be prepared to pay for a weather forecast: if he knows it will rain, he might do something else instead. Importantly, information only has value it improves the quality of your future decision.

When it comes to morality, many people have the view that studying ethics is useful in order to make them more confident about what they ought to do. But IB doesn't seem able to account for the possibility there is *value of moral information* (MacAskill, Bykvist and Ord, 2020, ch. 9). To see this, consider:

> ***Wise Philanthropist*:** You are 40% confident in theory A (on which e.g. AI safety is the priority) and 60% confident in theory B (on which e.g. poverty alleviation is the priority). You could decide your future allocation of resources now. Or you could study more moral philosophy. If you study more, you suspect you will end up 50% confident in each theory.

Recall that the sub-agents are certain in their theory and resources are allocated in proportion to the agent's credences. Given this, we can see that A would be wholly against studying - they stand only to lose - and B wholly for it - they stand only to gain.[8]

[7] As Greaves and Cotton-Barratt (2019) point out, in ordinary bargaining theory, a *disagreement point* represents what would happen if the bargaining parties cannot agree. However, when it comes to intra-personal bargaining, because the bargaining is only metaphorical, there is no clear empirical matter of fact about what the disagreement point would be.

[8] Patrick Kaczmarek makes the interesting suggestion that agents will also be motivated to acquire non-moral information that convinces the other agents to support their existing preferences; this is analogous to the way people will often seek information that will convince others to agree with them.

There are a couple of puzzles here. First, what would the agent do? Would A metaphorically overpower B and stop you from hitting the books? Or would A use the time it has allocated to pursue A's priority, whereas B would use some of its time allocation to study philosophy? As we've seen, bargaining is complicated.

Second, how can we capture the idea that we are able to gain moral information if all the sub-agents are already certain? Perhaps we should extend the metaphor by proposing the existence of some undecided sub-agents: the value of moral information relates to them forming a view.

## How grand are the bargains?

Bargaining gets more complicated the more we try to account for. Consider:

> ***Road Ahead*:** 40% credence in theory A on which reducing existential risk is the priority, and 60% in theory B on which the priority is alleviating poverty. You are a student and face two choices: you can opt for a career in either AI safety research or international development. You can also donate your spare resources to AI safety or poverty alleviation. On theory A, both your money and time are five times more valuable if put towards AI safety rather than poverty. On theory B, your resources towards poverty are twice as valuable.

We could take these choices in isolation in which case, naively, you split the pot with both your donations and your career, i.e. in the latter case you spend 60% of your career in international development and 40% in AI safety. However, splitting one's career seems impractical: if you try to split your career, you are unlikely to have much success in either field.

That suggests the sub-agents may well prefer to be able to negotiate both at the same time. In this case you might agree to specialise in one career but then donate considerably more to the other cause, a quite different outcome.

Greaves and Cotton-Barratt (2019) identify a challenge for IB: what we choose may depend on whether we consider a 'small world' (a simple set of options) or a more complicated 'grand world'. Whilst the maximally grand world is, in principle, the appropriate one to consider, this is often impractical. Hence, we should want our decision-theory to give approximately the same answer in the small worlds context as it would in the large one.

This doesn't strike me as a reason against, in principle using an IB approach to moral uncertaint, so much as a reminder that decision-making is complicated in practice. After all, even if we are morally certain, planning our future lives is already complicated; it doesn't seem to follow from that we should give up and follow simplistic rules.

## Fanaticism and the analogy to empirical uncertainty

Two further possible objections to IB are that it is insufficiently fanatical and it fails to be analogous to empirical uncertainty. I'll take these together and I suspect they are related.

MacAskill, Bykvist, and Ord (2020) respond to the accusation that MEC is objectionably fanatical by pointing out that fanaticism is equally a problem for expected utility theory in empirical uncertainty. If we are happy with expected utility theory - despite its fanatical results - then, by extension, we should be equally happy with MEC. So, fanaticism is an intuitive cost, but it's one we should expect to pay. Greaves and Cotton-Barratt (2019) similarly state it's unclear if fanaticism is objectionable or can be reasonably avoided.

A couple of quick replies. Firstly, most people seem to think that fanaticism is a problem, both for MEC and expected utility theory. Hence, insofar as one finds fanaticism problematic, it is an advantage for IB that it avoids fanaticism.

Second, if moral uncertainty and empirical uncertainty are analogous, then we should expect equivalent theories in each case. But, how confident should we be this is this the right analogy? Perhaps they are relevantly different. Here's a case that may motivate this. Consider two choices:

> ***Lives*:** You can pick (A) to save 1 life for certain, or (B) a 1 in a million chance of saving 1 billion lives.
>
> ***Life or Soul*:** You can pick (A) to save 1 life for certain, or (B) a 100% chance to convert someone to an unspecified religion. According to this religion, converting someone saves their soul and is equivalent to saving 1 billion lives. You assign a probability of 1 in a million that the claims of this religion are true.

It does not seem a mistake to choose (B) in *Lives*, but it does to pick (B) in *Life or Soul*. Yet MEC would treat these as structurally equivalent, with the latter choice being 1,000 times as choiceworth.

Intuitively, there is a difference between cases where there is (a) a low probability of a high payoff where that payoff certainly has value and (b) certainty of a payoff that you think is very likely to have no value.

Hence, there seems to be something especially troubling about fanaticism within moral uncertainty. IB is able to avoid this.

### What about regress?

Once we realise we are uncertain about morality, we face an apparent challenge of infinite regress: presumably we should be uncertain in our theory of moral uncertainty too. What would we do then?

I'm not sure if IB helps with this worry. IB makes sense of moral uncertainty by saying we need to distribute our resources to internal sub-agents who are certain in their view. It doesn't make sense to then ask what the sub-agents should do if *they* are uncertain: after all, we've stipulated that they are. However, when considering the option to moral uncertainty, your credence will still be somewhat split between IB, MFT, and MEC (and whatever else is on the table) as the ways to resolve moral uncertainty.

## 4. Taking stock of internal bargaining

Now we've got a sense of how IB would function in a range of scenarios and considered some problems, we can take more of a view of how plausible it is as an approach to moral uncertainty. I think most people would agree these are the desirable theoretical features, even if they disagree about their relative importance:

- Stake sensitivity (decisions can change if stakes change)
- Credence sensitivity (decisions can change if credences change)
- Does not require intertheoretical comparison of value
- Non-fanatical (decisions not dictated by low-credence, very high-stake theories)
- Robustness to individuation (decisions not changed by how moral theories are individuated)
- Provides a justification for 'worldview diversification' (beyond non-moral considerations e.g. diminishing marginal returns)
- Account for the value of *moral* information
- Avoids regress

Hence, if we put our three contender theories - MFT, MEC, and IB - side by side, it seems they would like *something* like this. I accept that I haven't explained all these fully here.

| | MFT | MEC | IB |
|---|---|---|---|
| Stake sensitivity | x | ✓ | ✓ |
| Credence sensitivity | ✓ | ✓ | ✓ |
| Intertheoretical value comparison unnecessary | ✓ | x | ✓ |
| Non-fanatical | ✓ | ✓ | ✓ |
| Robustness to individuation | x | ✓ | ✓ |
| Justifies worldview diversification | x | x | ✓ |
| Account for the value of moral information | ? | ✓ | ? |
| Avoids regress | x | ✓ | ? |

On this basis, internal bargaining looks like an appealing option and 'scores' better than the main alternative, MEC.

How much might it matter if we used IB rather than MEC? Potentially quite a lot. I've already noted in discussion of *Poverty or Robots*, MEC

# 5. A practical implication

What difference would it make in practice to adopt IB, rather than the alternatives of MFT or MEC? It's not easy to say much about this, given the substantial variety in what credences people place in different moral theories as we uncertainties is how the approaches to moral uncertainty function.

However, I do want to draw out and close on one practical implication we've noticed in passing already. One important recent idea with effective altruist thinking is *longtermism*, the view that improving the long-term future is the moral priority (Greaves and MacAskill, 2021). What's the case?

To quote Moorhouse (2021),

> *Three ideas come together to suggest this view. First, future people matter. Our lives surely matter just as much as those lived thousands of years ago — so why shouldn't the lives of people living thousands of years from now matter equally? Second, the future could be vast. Absent catastrophe, most people who will ever live have not yet been born. Third, our actions may predictably influence how well this long-term future goes. In sum, it may be our responsibility to ensure future generations get to survive and flourish.*

If longtermism is true, it would seem to imply that efforts to reduce existential risk, e.g. from unsafe AI are a higher priority than efforts to help people alive today, e.g. by reducing poverty.

At first glance, a key part of the argument is, speaking roughly, the moral priority we should attach to people alive today compared to future lives. After all, people alive today actually exist and we can make them better off. But future people may never exist, and it's puzzling to think that we really benefit them by creating them.

However, it would seem that, according MEC, these doubts are practically irrelevant: after all, surely we have *some* credence in a Total Utilitarian view and, on this view, we value present lives as much as merely possible future lives. Given *how many* lives there are could be in the future and the fact we can, presumably, affect the long-term, it seems we end up being pushed to longtermism even if we only a tiny credence in such a view (cf. Greaves and Ord, 2017). Therefore, we should abandon all our other, non-longtermist altruistic projects. This will strike many as objectionably fanatical.[9]

Saliently, internal bargaining does not seem to get this result. Instead, it seems the appropriate response is to engage in worldview diversification and split one's resources: you should commit your resources to longtermism in *proportion to the credence you have in moral views on which longtermism is true*, and the rest to non-longtermism causes.

I should stress two further points. First, IB would not prevent someone from committing (nearly) all their resources to longtermism if they have (nearly) all their credences in moral views on which it's true. IB merely avoids the (fanatical) result that everyone, almost no matter what their beliefs, should

[9] This conclusion may strike people as implausible because they don't think that longtermism is true *on any moral theory*: for instance, you could have 100% credence in Total Utiltiarianism, longtermism is false because we can't predictably and significantly influence the longterm. Such a person would not be objecting *specifically to fanaticism emerging from accepting MEC* - that is the objection I am concerned with here.

commit all their resources to it. Second, and conversely, IB implies that all agents should allocate *some* resources towards longtermism. Hence, IB in itself does not provide grounds to ignore longtermism altogether - longtermists may consider this a victory of sorts. I leave it to further work to consider exactly how internal bargaining may play out here.