Update: A more polished version is now online here.

Commenting on MSR, Part 2: Cooperation heuristics

This post assumes deep familiarity with the ideas discussed in Caspar Oesterheld's paper *Multiverse-wide cooperation via coordinated decision-making*. I wrote a short introduction to multiverse-wide cooperation in an earlier post (but I still recommend reading parts of Caspar's paper first, or at least this advanced introduction, because some of the things that follow build on topics that I did not cover in my introduction). So, in this post, I will comment on what I think might be interesting aspects of multiverse-wide cooperation via superrationality (abbreviation: MSR) and what I think might be its practical implications – if the idea works at all. I wrote this post to clarify my own thinking about MSR as well as to kick-start a wider discussion. I will focus particularly on aspects where I place more emphasis on certain considerations than Caspar does in his paper, though most of the issues I discuss are already noted by Caspar. A major theme of my comments will be exploring how the multiverse-wide compromise changes shape once we go from a formal, idealized conception of how to think about it, to real-world policy suggestions for humans. For the perhaps most interesting part of the post, skip to the section "How to trade, practically."

[*Epistemic status:* It is quite likely that my thinking in this post is misguided in important ways. Consider this to be a discussion starter rather than a confident statement about how one should think about MSR. Also note that I am outlining practical implications not because I am convinced that they are what we should do, but as an exercise in what I think would follow given certain assumptions. This can then help us zoom in on new MSR-related research questions.]

CONTENT

Decision heuristics for cooperation 2

Asymmetries amongst potential MSR participants 2

Asymmetries do not break MSR, but they make it messy 3

It is important to think in terms of cooperation heuristics 5

A failure mode to avoid 6

Inclusivity is not always better 9

Updateless compromise 10

How to trade, ideally 11

How to trade, practically 12

The relationship between “causal” cooperation and MSR 14

Some open research questions 15

Decision heuristics for cooperation

Under idealized conditions, each MSR participant would attempt to follow the same multiverse-wide compromise utility function (MCUF) reflecting the distribution of values amongst all superrational cooperators. In practice, trying to formalize a complete probability distribution over the MCUF, and consulting it for every decision, is much too noisy and effortful. A more practical strategy for implementing MSR is to come up with heuristics that approximate the MCUF reasonably well. Let’s call these *cooperation heuristics (CHs)*. An example for such a heuristic might be “Perform actions that benefit the value systems of other superrational cooperators considerably if you can do so at low cost, and refrain from hurting these value systems if you would only expect comparatively little gains from it.” This example heuristic is easy to follow and unlikely to go wrong. In fact, aside from the part about superrationality, it sounds like a great cooperation heuristic even for people who do not buy into MSR yet are interested in low-effort ways of cooperating with others for all the normal reasons (the ones that do not involve aliens). The main caveat about this particular CH is that it is very vague, and that the gains from trade it produces if everyone were to follow it are far from maximal. MSR may make it possible for us to give other value systems even more weight through further-reaching CHs, without thereby shooting ourselves into the foot.

Asymmetries amongst potential MSR participants

In the standard prisoner’s dilemma, the participants have symmetrical information and payoff structures. MSR is more messy: Superrational cooperators find themselves with information asymmetries, different goals, different biases and different resources at their disposal. Consider this non-exhaustive list of examples for potentially asymmetric features between MSR participants:

- **Frequency:** How common a value is amongst superrational reasoners.
- **Sunk costs, risk aversion:** The situation for proponents of different value systems may differ with regard to how much MSR would change their priorities. Potential MSR participants may therefore have differing levels of sunk costs, or different risk-reward tradeoffs, when they consider changing their priorities more towards the MCUF.
- **Cooperation saliency:** MSR considerations may be more salient to proponents of certain value systems than for others. (For instance, some people might be thinking about cooperation a lot, e.g. because their priorities pre-MSR are in part opposite to what others are pursuing, which makes it more likely that they will discover MSR early on and perhaps be more drawn towards making strong updates based on it.)
- **Knowledge about other value systems:** Value systems may differ in how much they know about other potential MSR participants. (For instance, there could be worlds where all evolved intelligent beings hold the same values.)
- **Degree of being known:** Value systems may also differ in how much *others* know about them: Simple/elegant value systems such as variants of utilitarianism are

presumably known and understood by many; whereas parochial value systems are known and understood only by few.

- **Benefitability:** Some value systems may allow easier ways of value creation than others, whose prioritization may be more complicated to get right.
- **Civilizational maturity:** Agents who reason about the nature of the MCUF may come to radically different conclusions regarding MSR depending on the stage of civilizational development they find themselves in. For instance, maybe later civilizations will contain more agents who reason about MSR.
- **Expertise:** MSR participants may differ greatly in the type of knowledge and expertise they have. Participants may often have comparative advantages for interventions favored by their own value system, simply because they may be more viscerally motivated for those interventions or may know more about the relevant prioritization implied by their value system.
- **Mainstream bias:** Minority value systems may be biased in favor of joining the things that are regarded as high status in the larger community. (Conversely, value systems that are attractive to contrarians may come with a bias against joining mainstream interventions.)

Asymmetries amongst potential MSR participants call into question whether we can indeed assume that our potential cooperators are finding themselves in sufficiently similar decision situations. To recap: We are assuming that MSR can work when two agents operate on highly similar decision algorithms and find themselves in highly similar decision situations. Under these conditions, certain approaches to decision theory, which I am for the sake of simplicity referring to with the umbrella term *superrationality*, recommend reasoning as though the decision outputs of the agents in question are logically entangled and output the same decisions. Asymmetric features amongst potential MSR participants now make it non-obvious whether we can still talk of the decision situations different agents find themselves in being “relevantly similar,” or whether they break the similarity because the conclusions that participants will come, whether to incorporate MSR into their behavior or not, are *affected by these differences.*

Asymmetries do not break MSR, but they make it messy

I think the answer depends on the level of abstraction at which the agents are looking at a decision, how they come to *construe* their decision problem. We can assume that agents interested in MSR will try to pick whichever **general process for selecting cooperation heuristics** that produces the largest gains from trade (under the assumption that everyone follows it).

Because the correct priorities of agents that pursue MSR may depend on asymmetric features between MSR participants, we can expect that “implementing MSR” could look quite different for different agents. Everyone tries to use cooperation heuristics that produce optimal benefits, but the cooperation heuristics would recommend different types of actions depending on whether an agent is in a particularly good situation for one action or the other.

To illustrate what I mean, consider agents who expect the highest returns from MSR from a focus on **convergent priorities** where they would work on interventions that are positive for their own value systems, but different from their top priority absent MSR considerations. This can be visualized as compromise clusters where a few different value systems mutually benefit each other through a shared priority. Value systems A, B, C and D may for instance have a shared priority x and value systems E, F, G and H may shared priority y. (By "priority," I mean an intervention such as "reducing existential risk" or "promoting consequentialist morality.") By focusing on convergent priorities, one only benefits a subset of all value systems in MSR *directly* (only the ones one shares such convergent priorities with). However, in theory we should now also expect there to be increased coordination between other value systems that form a cooperation cluster around *their* convergent priorities.

Similarly, following the (partial) CH of mostly cooperating with those value systems we **know best** (e.g. only with value systems we are familiar with from experience) makes it more likely that civilizations full of completely alien value systems will also only cooperate with the value systems they already know (which sounds reasonable).

Finally, whether MSR participants should spend all their resources on convergent priorities, or whether they should rather work on (other) **comparative advantages** they may have for greatly benefitting particular value systems depends on the specifics of the empirical circumstances the agents find themselves in. The tricky part about focusing on comparative advantages rather than (just) convergent priorities is that it might be one's comparative advantage to do something that is neutral or negative according to one's value system. In such a case, one needs to be particularly confident that MSR works well enough, and that one's CH is chosen/executed diligently enough, to generate sufficient benefits.

In practice, the whole picture of who benefits whom becomes quite complicated. A map of how different agents in MSR benefit each others' value systems would likely contain all of the following:

- Compromise clusters around convergent priorities: E.g. value systems [A, B, C, D] cluster around intervention x, and value systems [E, F, G, H] around intervention y.
- Partial overlap between some of these compromise clusters: E.g. value systems [A, F, H may share a common intervention z, on which they spend a non-zero percentage of their resources on.
- Arrows away from some of the convergent priorities that represent agents of value systems A – H focusing on personal comparative advantages some of the time, especially benefitting e.g. value systems [Q, P, R], which are particularly hard to benefit absent finding oneself with a comparative advantage for just that.
- Value systems that have no convergent priorities with other value systems and are only benefitted by considerations from comparative advantages (and also benefit others solely that way).
- Some arrows that are crossed out, representing interventions that proponents of a particular value system would pursue pre-MSR, but refrain from pursuing because they may hurt other MSR-participating value systems.
- …

The goal would be to pick one’s cooperation heuristic in whichever way that maximizes the gains from trade for all value systems, provided that all MSR participants pick their cooperation heuristic according to the same criteria. (If done properly and if the assumptions behind MSR are correct, this corresponds to maximizing the gains from trade for one’s own value system.)

Thinking in terms of cooperation heuristics is important

For figuring out what MSR implies for humans, I think it is important to think in terms of agents of moderate intelligence and rationality executing practical CHs, as opposed to ideal reasoners computing a maximally detailed MCUF for all decisions. Using heuristics corresponds to making a tradeoff between accuracy and practical concerns. Being accurate in one’s estimation of the MCUF and its practical implications for one’s own situation is obviously something important. If one makes too many mistakes, this lowers any gains from trade and may even result in net utility loss when comparing the outcome to never considering MSR in one’s actions in the first place. Particularly for value systems which are hard to benefit, accuracy in picking a sensitive enough cooperation heuristic is important, as the cooperation heuristic in question needs to guarantee that one notices when it is one’s multiverse-wide comparative advantage to benefit these value systems.

whether rare or hard-to-benefit value systems actually benefit from a cooperation heuristic depends on whether the heuristic is sensitive enough to notice situations where one’s comparative advantage is to benefit these value systems. This makes it challenging to pick simple heuristics that nevertheless react well to the ways in which all the features in decision situations can vary.

This is why I recommend being careful with talk such as the following:

“Intervention X [insert: global warming reduction, existential risk reduction, AI safety, etc] is good for MSR.”

To be clear, there is a sense in which this way of talking can be perfectly reasonable. From the perspective of the MCUF, majority-favored interventions receive a boost in how valuable they are as compared to evaluation from any single value system. Similarly, interventions that benefit value systems that are complicated to benefit also receive a boost if the MCUF incorporates variance normalization (see chapter 3 here for an introduction). This means, roughly, that one looks at the variance of how much value or disvalue is commonly at stake for each value system and compensates for value systems being hard to benefit. The reasoning is that one wants to incentivize all value systems to join the compromise. This may be especially important because it seems likely that for value systems that share few practical priorities, there may be other high-impact means for their proponents to benefit each other, such as proponents of one value system refraining from harming the other value system by reckless pursuit of their priorities.

If a value system is (for whatever structural reasons) particularly difficult to benefit, then for any of the rare instances where one is able to actually benefit said value system, it becomes important that those in the position to do so will in fact notice this and act accordingly.

All of the above paints a complicated picture. The worry is that some of these nuances will get lost in translation. When someone hears “Intervention x is positive for MSR,” they may do more of intervention x without ever checking what *other* interventions are positive too, and potentially *more* positive for their given situation. As soon as people start to take shortcuts, there is a danger that these shortcuts will disproportionately and *predictably* benefit some value systems and neglect others. (Shortcuts = cooperation heuristics produced by a dangerously low amount of careful thinking.) Even if everyone always does things that are positive according to the MCUF, it is possible for specific value systems to lose a lot of value in expectation or even suffer expected harm overall. The variance-voting / equal gains from trade MCUF is set up such that *if everyone tries to maximize it*, it distributes gains equally. There is no guarantee that if everyone just picks random stuff that is positive for the MCUF, this will be good for everyone. CHs have to be selected with the same principle in mind: We want to pick CHs which ensure equal gains from trade provided that everyone follows them diligently.

All of this suggests that whether an intervention being performed somewhere in the multiverse is “positive news” in expectation for all MSR participants is not only a feature of the intervention itself, but also depends on whether the executing agents have a sufficient comparative advantage for the intervention in question, or whether the intervention is chosen within a focus on convergent priorities. This suggests that in many contexts it might be epistemically safer to talk about CHs rather than concrete interventions being what is “positive for MSR.”

A failure mode to avoid

Asymmetries amongst MSR participants and the issue with choosing CHs in a way that distributes the gains from trade symmetrically make it tricky to pick cooperation heuristics wisely. I am particularly concerned about the following failure mode:

Superrationalizing: When the CH you think you follow is different from the CH that actually guides your behavior.

For instance, you might think your CH produces the largest expected gains given practical concerns, but, unbeknownst to you, you only chose it the way you did because of asymmetric features that give you a disproportionate benefit. Others, who you thought will arrive at the same CH, will then adopt a different CH than the one you think you are following. You therefore lose out on the gains from trade you thought your CH would produce.

Similarly, one might think one is following a CH that produces large gains from trade for one’s value system, but if the de facto execution of the CH one thinks one follows is too sloppy, then one has no guarantee for the predicted gains from trade to materialize.

For better illustration, I am going to list some examples for different kinds of superrationalizing in a more concrete context. For this, let me first introduce two hypothetical value systems of MSR participants: *Straightforwardism* and *Complicatedism*.

Straightforwardists have practical priorities that are largely shared by the majority of value systems interested in MSR. Proponents of Complicatedism on the other hand are not excited about the canon of majority-favored interventions.

This is our setup. Now, for one instance of superrationalizing, let us assume that the Straightforwardists pick their CH according to the following, implicit reasoning: “When MSR participants reason very crudely about the MCUF and only draw the most salient implications with a very simple CH, such as looking for things that benefit a lot of value systems, this will be greatly beneficial for us. Therefore, we do not have to think too much about the specifics of the MCUF and can just focus on what is beneficial for many value systems.”

By contrast, proponents of Complicatedism may worry about getting skipped in the compromise if people only perform the most salient, majority-favored interventions. So they might adopt a policy of paying extra careful attention to value systems never getting harmed by MSR in expectation, and therefore focus their own efforts disproportionately on benefitting the value system *Supercomplicatedism*, which only has few proponents and whose prioritization is very difficult to take into account.

Of course, MSR does not work that way, and the proponents of the two value systems above are making a mistake by, perhaps unconsciously/unthinkingly, assuming that other MSR participants will be affected symmetrically by features that are specific to only their own situation. The mistake is that if one pays extra careful attention to value systems never getting harmed by MSR because one’s own value system is in a minority that seems more at risk than the average value system, then the reasoning process at work is not “No matter the circumstances, be extra careful about value systems getting harmed.” Instead, the proper description of what is going on then would be that one unfairly privileges features that are only important for one’s own value system. To put it differently, if proponents of Straightforwardism think “I allow myself to reason crudely about MSR partners, therefore other agents are likely to think crudely about it, too – which is good for me!” they are missing that the reason *they* were tempted to think crudely is not shared by all other agents in the compromise.

In order to maximize the gains from trade, proponents of both value systems, Straightforwardism and Complicatedism, would have to make sure that they use a decision procedure that, in expectation, benefits them (in proportion to how prevalent and powerful the proponents are) for every instance where it is being applied. Straightforwardists have reason to pick a CH that also helps Complicatedists sufficiently much, and Complicatedists are incentivized to not be overly cautious and risk averse. If implemented properly, asymmetries between potential MSR participants *cannot be used to gain an unfair advantage*.

Now, for a slightly different example of superrationalizing, consider a case where Complicatedists naively place *too much faith* into the diligence of the Straightforwardists. They may reason as follows:

> “The majority of compromise participants benefit from intervention Z. Even though intervention Z is slightly negative or at best neutral for my own values, I should perform intervention Z. This is because if I am diligent enough to support Z for the common good, as it seems best for a majority of compromise participants and therefore an obvious low-hanging fruit for doing my part in maximizing the MCUF, other agents will also be diligent in the way they implement MSR. Others being diligent then implies that whichever agents are in the best position to reward my own value system will indeed do so.”

This reasoning is sound in theory. But it is also risky. Whether the Complicatedists reap gains from trade, or whether the true decision procedure they follow (as opposed to the decision procedure they *think* they follow) implies that they are shooting themselves in the foot, depends on their own level of diligence in picking MSR implications. The Complicatedists have to, through the CH they *de facto* follow, ensure that the agents who are in fact in the best position to help Complicatedism will notice this and act accordingly.

It seems to me that, if the Complicatedists put all their resources into intervention Z and never spend attention researching whether they themselves might be in a particularly good position to help rare value systems or value systems whose prioritization is particularly complicated, then the reasoning process they are de facto following is itself not as diligent as they require their superrational cooperators to be. If even the Complicatedists (who themselves do not benefit from the majority-favored interventions) end up working on the majority-favored interventions because they seem like the easiest thing to pick out, why would one expect agents who actually benefit from this “low-hanging fruit” to ever work on anything else? The Complicatedists have to make sure that they work on majority-favored interventions if and only if it is actually their comparative advantage to do so. This may be difficult to ensure, because one might expect that people start rationalizing, especially when majority-favored interventions tend to be associated with high status, or tend to draw in Complicatedists high in agreeableness who are bothered by lack of convergence in people’s prioritization.

For allocating personal comparative advantages in the way that produces the greatest gains from trade, one has to find the right mix between exploration and exploitation. It is plausible that MSR participants should *often* focus on majority-favored interventions, because after all, the fact that they are *majority*-favored means that they make up a large portion of the MCUF. But next to that, everyone should also be on the lookout for special opportunities to benefit value systems with idiosyncratic priorities. This should happen especially often for value systems that are well-represented in the MCUF, but maybe one could also make use of some randomization procedure to sometimes even spend time exploring the prioritization of comparatively rare value systems (see also the proposal in “How to trade, practically”).

Regarding the use of randomization procedures, it should be noted that it can be difficult to properly commit to doing something that may cost social capital or is difficult to follow through with for other reasons. Illusory commitments weaken or even destroy the gains from trade one in expectation receives through this aspect of MSR. Proper introspection and very high levels of rationality become important for not shooting oneself into the foot when attempting to get MSR implications right.

An intuition I got from writing this section is that it tentatively seems to me that cooperation heuristics that exploit convergent priorities (in particular when the resulting intervention benefits one's own value system) are less *risky* (in the sense of it being harder to mess things up through superrationalizing) than trades based on comparative advantages. However, because people are more likely to be able to coordinate mutual focus on convergent priorities even through ordinary means of cooperation, cooperation heuristics that emphasize considerations of personal comparative advantages are likely to produce particularly high gains from trade.

Inclusivity is not always better

Which value systems in particular MSR participants should benefit depends on their situations and especially their comparative advantages. I have advocated for the idea that we should limit our cooperation heuristics to considering value systems we know well.

One might be tempted to assume that this would be a bad thing, as limiting how inclusive one is with benefitting value systems different from one's own determines how many value systems will be incentivized to join our compromise in total. So perhaps low inclusivity in this way means that one's decisions now only influence a smaller number (or a lower but still infinite kind of *density*) of agents in the multiverse. However, it is important to note that MSR never manages to bring other agents to follow one's own priorities exclusively; it only grants you a *proportionate share* of the attention and resources of some other agents. The more types of compromise participants are added to a cooperation heuristic, the smaller said share of extra attention one receives per participant. (Consider: If I have to think about what my comparative advantage is amongst three value systems, that takes less overhead than figuring out one's comparative advantage amongst three hundred value systems.) This means that there is no overriding incentive to choose maximally inclusive cooperation heuristics, i.e. ones that in expectation benefit maximally many value systems of superrationalists in the multiverse.

Note that this means that one cannot make a strong wager in favor of MSR of the sort that, if MSR works, our decisions have a vastly wider scope than if it does not work. While it is true that our decisions have a wider scope if MSR works, this is counterbalanced by us having to devote attention to different value systems in order to make it work. MSR's gains from trade do not come from the large total numbers of participants, but from exploiting convergent priorities and comparative advantages. So while it is not important to consider maximally many plausible value systems in one's compromise, it *is* important that we do include whichever value systems we expect large gains from trade from (as this superrationally ensures that others follow similarly high-impactful cooperation heuristics).

If one had infinite computing power and could at any point distill the implications of an ideal MCUF that contains all agents interested in MSR, then a maximally inclusive compromise would give the highest benefits. However, given that thinking about the prioritization of other value systems (especially obscure ones that only make up a tiny portion of the MCUF) comes with a cost, it may not be worthwhile to invest resources into ever more sophisticated cooperation heuristics solely with the goal of making sure that we do not forget value systems we could in theory benefit. This reasoning supports the intuition that the best way to draw implications from MSR is by cooperating with proponents of value systems that one already causally interacts with, because these are the value systems we know best and are therefore in a particularly good (and also non-arbitrary) position to benefit.

Updateless compromise

So far, I have been assuming that agents only follow cooperation heuristics that, at the stage of execution, the agent believes will generate positive utility according to their own value system. This sounds like a reasonable assumption, but there is actually a case to be made for exceptions to it. I am talking about updateless versions of compromise.

Suppose I am eating dinner with my brother and we have to agree on a fair way of dividing one pizza. Ordinarily, the fair way to divide the pizza is to give each person one half. However, suppose I like pizza a lot more than my brother does, and that I am also much more hungry. Here, we might have the intuition that, whether person A or person B likes the pizza in question more, or is more hungry on that specific occasion, was a matter of chance that could just as well have gone one way or the other. Sure, one brother was born with genes that favor the taste of pizza more (or experienced things in life that led him to develop such a taste), but there is a sense in which it could also have gone the other way round. Updatelessness is the idea that extra knowledge should never harm the potential for gains from compromise. With this in mind, it could mean that my brother and I should disregard (= “choose not to update”) on knowledge that one *specific/known* person now has the less fortunate pizza preferences in the specific instance we are in, because there were points in the past where we could have agreed on a method for future compromise on things such as pizza eating, which in expectation does better than just dividing goods equally. Not knowing whether we ourselves will be hungrier or less hungry, it seems rational to commit to a compromise where the hungrier person receives more food. (There is also a more contested, even stronger sense of updatelessness that is not based on pre-commitments.)

Updatelessness applied to MSR would mean to optimize for a MCUF where variance normalization is not applied on all the things we currently know about the strategic position for proponents of different value systems, but instead to a hypothetical “point of precommitment.” Depending on the version of updateless at play, this could be the point in time where someone started to understand decision theory well enough to consider the benefits of updatelessness, or it could even mean going back to the “logical prior” over how much different value systems can or cannot be benefitted. (I do not understand much about either logical priors or how to distinguish different versions of updatelessness, so I will just leave it at that and hope that others may do some more thinking here.)

The inspiration for updateless compromise is that the gains in case one ends up being on the lucky side weigh more than the losses in case where one is not. Maybe it is not apparent from the start which value system corresponds more to something like Complicatedism or something like Straightforwardism, and the sides could in theory also be reversed in some world-situations across the multiverse, depending on the things that happen in the worlds in question. There is therefore a case to be made for committing towards updateless compromise before thinking more about MSR implications in further detail. (Or more generally, there is a case to be made for precommitting towards updalessness in all future decision-situations where this has benefits given the knowledge at the time of precommitment.)

While I think the arguments for updatelessness are intriguing, I am skeptical whether humans can and should try to trick their brains into reasoning completely in updateless terms. And I am even more skeptical about using updateless compromise for MSR in particular.

Next to the psychological difficulties with updatelessness and worries whether humans are even capable of following through with the implications after learning that one is on the losing end of a compromise, another problem with updateless MSR is also the apparent lack of a true original position (besides the extreme view where one just goes with a logical prior). We have previously discussed asymmetric features amongst potential MSR participants. Even someone who has not given much thought to the relative prioritization of different value systems will probably have a rough idea whether their value system is more likely to benefit from updateless compromise or not. Even small asymmetries can break the entanglement of decision algorithms: If I commit to be updateless because I have a good feeling about being on the winning side, I cannot expect other agents who may not share said feeling to commit as well.

Having said all that, I guess it might be reasonable though to already commit to having precommitted to be updateless in case that, after thinking more about the merits and drawbacks of the idea, one concludes that a past commitment would in fact have been the rational thing to do. (Though it is plausible that I am just being confused here.)

How to trade, ideally

Without (strong versions of) updatelessness, the way we ensure that our actions lead to MSR benefits is to diligently follow cooperation heuristics that do not disproportionally favor our own values. (Otherwise we would have to conclude that others are disproportionally benefit *their* values, which defeats the purpose.) This means that, in expectation, all the value systems should receive a substantial portion of attention somewhere in the multiverse. Ideally, assuming there were no time or resource constraints to computing a compromise strategy, an ideal reasoner would execute something like the following strategy:

1) Set up a weighted sum of the utility functions of superrationalists in the multiverse.

2) Set the weights such that when universally adopted, everyone gets the same expected gains from compromise (perhaps relative to the agents' power).

3) Maximize that utility function.

The way to coordinate for each value system to have resources allocated to its priorities is to maximally incorporate comparative advantages in terms of expertise and the strategic situation of the participating agents. Step 2) in the algorithm above is therefore very complicated, because it requires thinking about all the ways in which situations across the multiverse differ, where agents are in an especially good position to benefit certain value systems, and how likely they would be to notice this and comply. To illustrate this complexity, we can break down step 2) into further steps. Note that the following only gives an approximate rather than exact way to solve the problem, because a proper formalization for how to solve step 2) would twist knots into my brain.

2.1) Outline the value systems of all superrationalists and explore strategic prioritization for each value system in all world situations to come up with a ranking of promising interventions per world situation per value system.

2.2) Adjust all these interventions according to empirical compromise considerations where one can get more value out of a given intervention by tweaking it in certain ways: For instance, If two or more value systems would all agree to change each other's promising interventions to different packages of compromise interventions that are overall preferable, make the change.

2.3) Construct a preliminary multiverse-wide compromise utility function (pMCUF) that represents value systems weighted according to how prevalent they are amongst superrationalists, and how influential its proponents are.

[If compromise was updateless, then perhaps that pMCUF would already be all that is needed.]

2.4) Compare the world situations of all participants in MSR, predict which interventions from 2.2) will be taken by these agents who are approximating the pMCUF, and calculate the total utility this generates for each value system in the preliminary compromise.

2.5) Adjust the weights in the pMCUF with a fair bargaining solution in such a way that all value systems will get sufficiently many benefits that they are properly incentivized to join the compromise. This eventually gives you (a crude version of) the final MCUF to use.

[Step 2.5 ensures that value systems that are hard to benefit also end up receiving some attention. Without this step, hard-to-benefit value systems would often end up neglected, because MSR participants would solely be on the lookout for options to create the most total value per value system, which disproportionately favors benefitting value systems that are easy to benefit.]

How to trade, practically

Needless to say, the analysis above is much too impractical for humans to even attempt to approximate with steps of the same structure. In order to produce actionable compromise plans, we have to come up with a simpler proposal. In the following, I'll try to come up with a practical proposal that, if anything, tries to err on the side of being *too simple*. The idea being that if the practical proposal below seems promising, we gain confidence that implementing MSR in a way that incentivizes sufficiently many other potential participants to join is realistically feasible. Here the proposal, in very sketchy terms:

1) Only include value systems in the MCUF that we can observe on earth. Preliminarily weight these value systems according to how many proponents of said value system interested in MSR there are, and how influential these proponents are.
2) *Flatten* the distribution of value systems from 1) based on prior expectations of how represented a value system would be without founder effects or general path dependencies. This could be done based on survey data on people's intuitions about common value systems (perhaps also including the intuitions of people who are skeptical about superrationality).
3) Figure out which interventions are particularly valuable for each value system, e.g. by communicating with its proponents and checking their reasoning if you think you might be better than them at drawing correct implications.
4) For value systems whose prioritization one is not very familiar with, randomize and only spend time exploring their prioritization with some non-zero probability that seems appropriate (i.e., is sensible from an all-things-considered exploration vs. exploitation tradeoff and is proportional to how prevalent the value system is). Note that, if the randomization procedure made you explore the priorities of a particularly rare value system, you now are much more likely to have a comparative advantage at benefitting it.
5) Adjust the interventions from above to make them more "positive-sum:" Deprioritize interventions where proponents of different value systems would be harming each other; adjust interventions to make them more beneficial for other value systems if the cost is low enough; adjust interventions to make them less harmful for other value systems if the cost is low enough; highlight interventions that are positive for many different value systems, etc.
6) Think of interventions that are good for everyone (or most value systems at least) but not good enough to make it on anyone's list.
7) Get a rough sense, perhaps just intuition or based on some quick calculations, on which value systems lose a disproportionate amount of value in step 5), and take a note to give them extra weight. Also give extra weight to interventions that are positive for many different value systems as identified in step 5).
8) Think about your competitive advantages as compared to other proponents of MSR at following all the (adjusted) interventions you got out of the previous steps. If one thing clearly sticks out as your comparative advantage amongst people interested in MSR, focus largely on that. If multiple interventions might plausibly be your

comparative advantage, use randomization with weights that represent the weights from steps 2) plus adjustments in step 6).

9) Keep an eye out for low-effort ways to benefit value systems other than the ones you're currently focusing on. Perhaps even institutionalize thinking about this by e.g. scheduling time every month where you randomly receive one value system and spend one hour thinking about how to benefit it, and if you have an idea that seems promising enough and is not prohibitively costly, follow through implementing it. (The idea is that his heuristic makes use of sharply diminishing returns for low-effort interventions.)
10) If possible, coordinate with other proponents of MSR to allocate resources in a better-coordinated fashion. Make use of gains from scale by gathering people interested in MSR and generally "strong cooperation," rank them in terms of various comparative advantages, and coordinate who focuses on which interventions (this makes it much easier to figure out what one's comparative advantages are).
11) Sanity check: Go through the heuristics Caspar lists in his MSR paper to see whether the procedure you are set on following has somehow led you to something crazy.

Note that point 5) also includes very general or "meta" interventions such as encouraging people who have not made up their minds on ethical questions to simply follow MSR rather than waste time with ethical deliberation.

Admittedly, the above proposal is vague in many of the steps and things often boil down to intuition-based judgment calls, which generates a lot of room for biases to creep in.

However, if people genuinely try to implement a cooperation heuristic that is impartially best for the compromise overall, then biases that creep in should be equally likely to give too much or too little weight to any given value system.

The relationship between "causal" cooperation and MSR

Causal interaction and cooperation with proponents of other value systems who are also interested in MSR can be highly useful as part of a cooperation heuristic, but one does not have to think of these other people as "actual" MSR compromise partners. It is debatable whether one's own decision-making is likely to be relevantly logically entangled with the decision-making of some humans we causally interact with. Whether this is the case or not, MSR does not require it. Besides, even if such entanglement was likely, the possibility of checking up on whether others are in fact reciprocating the compromise may break the entanglement of decision algorithms (cf. the EDT slogan "ignorance is evidential power").[1]

So the idea behind focusing on cooperating with the proponents of value system that we know and can interact with is not that we are superrationally ensuring that no one defects in causal interactions. Rather, the idea is that, if MSR works, each party has rational reason to act as though they are correlated with agents in other parts of the multiverse, where

[1] Decision theories incorporating updatelessness would continue to cooperate even after observing the other party's decision, if the reasons from similarity of decision algorithms were strong enough initially.

defection in expectation hurts their own values. This is what ensures that there are no incentives to defect. If one were to defect, one may gain an unfair advantage locally in casual interactions with others, yet one loses all the benefits from MSR for other parts of the multiverse.

Note that this leaves the problem that agents can fake to epistemically buy into MSR even though they may be highly skeptical of the idea. If one is confident that MSR will never work, one may be incentivized to lie about it and fake excitement. (Though I think this sounds like a terrible idea for the epistemic damage it would do to the community and all the non-MSR arguments against naive consequentialism.)

Some open research questions

To figure out whether we can trust the reasoning behind MSR, there are many things to potentially look into in more detail. Personally, I am particularly interested in the following questions:

- **Underlying assumptions:** Is MSR based on the correct decision theoretical assumptions? What exactly are the things we want to be “relevantly similar” between us and other agents elsewhere in the multiverse for MSR to work? Are values and decision procedures distributed orthogonally among agents in the multiverse? Or does the overwhelming majority of copies of my own decision algorithm share my specific values? I expect progress on these questions to come from naturalized induction and decision theory, and from getting a better idea on how one should think about the multiverse (or different multiverse proposals).
- **Estimating the gains from trade:** Whether it is warranted to change one’s actions to incorporate MSR considerations depends on both one’s credence in the underlying assumptions behind MSR being correct, and on the expected gains from trade in case the assumptions are correct (as well as the losses if not). It would be very valuable for people to think more about how large the potential gains from compromise would be for a well-executed cooperation heuristic such as “How to trade, practically.” Perhaps there are useful examples in the economics literature?
- **How to think about comparative advantages:** How can we estimate whether a perceived comparative advantage for helping a given value system is strong enough or not? Are there prudential reasons for favoring erring on the side of conservatism vs. experimenting with comparative advantages, or maybe the other way around? Perhaps some of the research on portfolio approaches to global prioritization could be informative for MSR as well.

Acknowledgments

Thanks to Caspar Oesterheld, Johannes Treutlein, Max Daniel, Tobias Baumann and David Althaus for comments and discussions that helped inform my thinking. I first heard the term “superrationalizing” in the context of superrationality from Carl Shulman.