Rationally Speaking #230: Kelsey Piper on “Big picture journalism: covering the topics that matter in the long run”

Julia: Welcome to Rationally Speaking, the podcast where we explore the borderlands between reason and nonsense. I'm your host Julia Galef, and my guest today is Kelsey Piper.

Kelsey is, in my opinion, one of the best new journalists out there. She writes full time for Future Perfect, which is a branch of Vox that's devoted to topics that have the largest impact on the world -- as opposed to just sort of covering topics that are new, as in news.

Kelsey has also been a blogger for years, which is how I started following her. Her Tumblr, “The Unit of Caring,” is one of my favorite things to read. So we're gonna start by talking about some of the work she's been doing for Future Perfect. And then transition into talking about some of her personal writing on topics like morality and mental health.

So Kelsey, welcome, thank you so much for being here.

Kelsey: Thanks so much Julia.

Julia: Why don't you tell our listeners a little bit about how Future Perfect came to be, what's its origin story?

Kelsey: Yeah, so Future Perfect is funded by the Rockefeller Foundation, and my understanding is that they were interested in the way that having an outlet where some people could focus full time on a question, and on coverage of that question, sort of influences the broader society. So for example how outlets like Breitbart came to be, and had a relentless focus on some conservative issues and sort of brought those more into the mainstream. And they were very interested in what would it look like to do this for sort of friendly altruistic cosmopolitan centrism.

And when they talked to Vox about this -- Vox has of course Ezra Klein and Dylan Matthews, who are interested in effective altruism. And they thought effective altruism is sort of a good inspiration and like, grounding source for a project like this.

So Future Perfect is effective altruism inspired, and draws a lot on that, and I think that's definitely a big part of our target audience. Although it's not an effective altruist outlet, and it covers a lot of issues that don't come up in effective altruism.

Julia: What would you say is in that area of non overlap? What's something that would count as an important topic impacting the world but wouldn't fall under the umbrella of EA?

Kelsey: So EAs have, I think rightly, been very wary of getting too involved in politics. But that is obviously an important topic for the world. And maybe the marginal impact of one person is very small, so it makes sense for effective altruism not to focus there.

But it's still where a lot of world-affecting decisions happen. So Future Perfect does consider it within our purview, and will cover for example the anti-poverty plans of 2020 candidates, or which redistribution schemes look like they would work the best, or sort of questions like that.

Which, I think it makes a lot of sense for effective altruism to sort of not focus on, given how contentious they can be and how many urgent neglected priorities there are. But which are a pretty natural fit for what Future Perfect is doing.

Julia: Yeah. The thing that's just now occurring to me -- tell me if this sounds right -- is that discussions in the world of EA are kind of filtered by intervention, so they're all about "What could we, a group or an individual, do now, on the margin, to have a large positive impact?" Whereas the coverage in Future Perfect is partly about that and partly about just prediction. Like, "Even if we couldn't affect which policy gets passed, it is still interesting and important to discuss what the likely effects of this or that policy are going to be." If they're you know, plausibly large.

So is that distinction, between intervention versus prediction, does that seem right to you?

Kelsey: Yeah, that seems like a great articulation of the distinction. Like Future Perfect is more of a ... Vox's background is "explain the news," right? So Future Perfect is doing a lot of explaining big topics, even if they're big topics where there's no obvious opportunities for an individual to act on the margin. And maybe it's not so central for EA.

Julia: Right. Cool, okay. Well, let's talk about one of the most recent articles you wrote. So I just saw ... You know, I'm constantly on Twitter, and I just saw that you were having a friendly disagreement on Twitter with another former podcast guest of mine, Rob Reich, who's a Stanford professor who wrote the book Just Giving. And it was sparked by one of your recent articles, which was about how the now infamous Sackler family, which arguably played a major role in causing the opioid crisis we're suffering from, through their ownership of Purdue Pharma, how the Sackler family is having some of their philanthropic donations now refused by their recipients. I guess mainly the museums they donate to.

Kelsey: Yeah.

Julia: In the article you examine the question of “Well, should charities refuse donations from people who may not have acquired their fortune in maximally ethical ways?” And I think Rob was more on the side of “Yes, they should refuse the donations,” and you were more on the side of “Often no, and it really depends.”

So do you feel like you understood the crux of your disagreement?

Kelsey: Yeah, I think so. So Rob thinks in terms of justice a lot more strongly than I do. I think effective altruists often come at the world from a very utilitarian perspective, “Where is the money gonna do the most good? That's where the money should be.” And I do believe that has a lot of value.

I think Rob's point was: You know, to some extent if we're making money off things as unethical as marketing tobacco in developing countries, or causing the opioid crisis by giving misleading instructions about how to dose opioids, you know, there's this justice consideration, that to his mind overrides “But where does the money do the most good” considerations.

Julia: Yeah.

Kelsey: And I think also, he thinks that we aren't really choosing between “they donate to charity” or “they keep the money.” From that perspective, of course they should donate to charity; do we want them to keep the money?

But his point was: Maybe if we're critical of this in the right way, we can get them to pay the money back to the people harmed, which is more just, as an outcome.

Julia: Does that seem plausible to you?

Kelsey: So in the case of the Sacklers, I think a lot of what I ended up being curious about and being troubled by was how much we can declare them an exception to our ordinary norms here, and how much they're more on the end of a spectrum.

Julia: A bleak take.

Kelsey: Yeah, so from one perspective, more billionaires make money in a good way -- by creating products of immense value which people are willing to pay them lots of money for. And we're talking about what to do with the very small segment of billionaires who are not like that, and make money unethically.

And you know, from another perspective, which is one I hear articulated a lot at Vox in particular, is billionaires are mostly doing some shady stuff. Especially when startups are getting off the ground, a lot of them are breaking laws and cutting corners and ignoring data privacy. And none of them are really in this commendable category where they made their billion dollars. Maybe JK Rowling. But I think there are some people that think she's pretty much the only billionaire who just became a billionaire by making something people wanted.

Julia: And she's losing a lot of good will now with her retroactive ... Changing of the canon. Or you know, adding a bunch of details, ten years after the fact, to her books.

Kelsey: Yeah. So to a lot of the world, I think there are no good billionaires. And if you think there are no good billionaires, then the question of "what do we do about bad billionaires" ... I have a hard time preferring the justice answers because I do ultimately want money to go where it's needed the most.

If you think that there are lots of good billionaires, which is more the side I come down on -- I think Jeff Bezos is rich because he made the world a much, much better place and collected some of the value from doing that, right? From that perspective, then I think it seems fine to focus on justice in our handling of the Sacklers. Because they're not our primary mechanism by which philanthropy gets funded and important things get money. They're kind of a little bit of a sideshow. And saying we're going to handle that by settling the wrong that they did and not by trying to distribute their money optimally, it doesn't give me nearly as much pause.

Julia: That is really interesting, because I had sort of assumed that the crux between you two was, like, justice versus utilitarianism. Or possibly this empirical crux of "Is it even plausible that if we refuse to let these rich families donate their money, that we could instead get the retributive justice outcome, that maybe is the best?" And that you know, you and Rob disagreed on how plausible that outcome was.

But it hadn't occurred to me that the crux might just be, what percentage of billionaires would we consider bad? Because it might just make sense to use very different rules in those two different worlds.

Kelsey: Yeah, and I do think the other things you mentioned are also disagreements. But definitely a lot of what made me sort of hesitate about the Sackler case was: if I articulate a principle here, and then someone goes "You know, Jeff Bezos doesn't pay warehouse workers very well, and this principle of yours should apply to him too," how am I gonna feel about that?

Julia: Right. Yeah. I've become ... So in conversation with Rob partly, I've become a little more sympathetic now than I used to be to the argument that we should care about whether philanthropists are receiving status, and respectability, and legitimacy, for their donations. It used to just seem like a red herring to me. Or basically, my view four or five years ago or something was: “On the margin, our choice is between someone gives the money to a good cause, or even like a mildly good cause, or they keep it for themselves, how could you possibly be against them giving the money away? That's crazy.”

And all these people who got mad when Jeff Bezos or Mark Zuckerberg pledged to give some of their fortune away -- I sort of thought they were focused on the red herring of “Well, I don't want anyone to say anything good about this person who I'm mad at for these other, maybe legitimate, reasons. So I have to oppose everything they do, instead of saying that this thing is good.” So I kind of thought it was confused.

Now, I'm more sympathetic to the idea that these issues are all kind of bound up and they're really hard to separate. And so if someone is going to acquire a lot of status and respectability in society for their donations, there may not be a way to separate that, from “Yes, they're doing these bad things but also it's good that they give their money away.” We might not be able to have those two totally separately.

That's my best steel man of the Rob Reich case, I don't know if he would endorse it himself though.

Kelsey: Yeah, parts of that definitely resonate with me. I think I end up listening to some of the things Rob says and going, “Yeah, so we want a balancing test, where we consider how much good they're doing and how much status they're getting.”

Julia: Right, right.

Kelsey: It seems like that's not quite what Rob believes, like I would [crosstalk 00:12:40].

Julia: That's not a very justice ...

Kelsey: Yeah, yeah.

Julia: Like if you're kind of a deontologist, like some things are right to do and some things aren't, and it's not about like measuring the consequences. Once you start talking about balance and measuring, you're back in utilitarian land.

Kelsey: Yeah. And I think I'm able to meet Rob as far as “Yeah, that's a consideration to balance against the other considerations.”

Julia: Right, yeah. Well put. Cool.

Let's move on to one of your biggest and most influential pieces. It's titled "The Case for Taking AI Seriously As a Threat to Humanity." And you've actually written two versions of this, the full version -- which even itself, at ... How long was it? Was it 5000 words?

Kelsey: It's like 6000, yeah around that.

Julia: Okay yeah. So even that had to simplify and abbreviate a lot.

Kelsey: Absolutely. Yeah, I absolutely cut a lot there.

Julia: Right. I mean I'm not faulting you at all, it's just you know, you can't write a book in an article.

And then you also had more recently a super abbreviated version that's like 500 words, which I was pretty impressed you were able to get it down another order of magnitude.

So, it seems like you approach this piece from the perspective of identifying and addressing the main objections to, or confusions about, the AI risk thesis. What emerged as the main objections that people have, to the idea that we should take AI seriously as a threat?

Kelsey: So I think there's a couple of them. One is people have a hard time imagining a concrete scenario by which something happening on a computer is dangerous, that doesn't sound you know, completely absurd.

Julia: Yeah.

Kelsey: And to some extent the AI community's in a little bit of a bind here. Because a lot of people have said "I don't want to describe a specific scenario that I think is actually quite unlikely. That seems dishonest to me, to describe something that I don't think is how it's actually gonna happen."

Julia: Or could even potentially set them up for criticism, like, "Oh so you're telling this specific story, this is kind of like science fiction -- you think you know what's gonna happen?"

Kelsey: Yeah. Yeah. But then if you don't tell stories and you're just like -- like Jaan Tallin, who I was talking to about this recently, was like, "Well, what I say is, you know, if cockroaches are trying to imagine how humans will kill them, they might imagine cockroaches with lasers attached to the front or something. They probably won't imagine spray, because they just don't have any of the concepts to come up with spray. And it's like that."

Julia: I'm just laughing at the idea of a cockroach imagining lasers attached to cockroaches, like ... Anyway, go on.

Kelsey: Swords attached to cockroaches, I don't know.

The balance I've ended up striking is describing a couple of ways that I, if I were on a computer and thought very fast and had a lot of money, could end the world. And then going “Yeah, it's probably gonna be more complicated than that, but you know, since at minimum you could do that, I'm pretty scared.”

Julia: Right.

Kelsey: And that's been kind of the best balance I've found, between being honest about the fact that we don't know, and it's not gonna look like any neat scenario we can come up with, while also giving people a concrete idea that yes, if you are in fact just on a computer, think really fast and have a lot of money, there are some ways that you could do a ton of harm.

Julia: Yeah. That's really helpful actually. And that's been a sticking point for me, the whole time that I've been engaging with the AI risk argument and community -- is just feeling caught between, well, the abstract argument is too abstract for me to really feel like I can get a handle on, or know how to take seriously. And any specific scenario feels too implausible. And so I don't really know how to engage with this.

Kelsey: Yeah.

Julia: So I find the compromise pretty helpful. Were there other objections?

Kelsey: So a big part of this piece was a result of me talking with the people at Vox, who are mostly effective altruism oriented, mostly very smart and informed, and pretty skeptical going in, of AI risk. And sort of saying, what are your questions?

And one that comes up a lot is, “Is this really a bigger deal than climate change? Like, where should this be on our list of priorities?” You know, there's a lot of things that seem like they might menace humanity making it through the 21st century intact.

And that's a hard one to do justice to, you know, without sounding dismissive of other concerns or anything. But there I tend to just come down on: There are like fewer than 50 people working full-time on existential risk for general AI. You know, a decade ago it was worse than that, and there were probably fewer than ten. That seems like too few. It should be a couple hundred. We don't need to take a stance on where this ranks among other global priorities to reach that conclusion, necessarily.

Julia: Yeah. This is a general pattern that I keep noticing. I talked about it a bit on a recent episode with Rob Wiblin from 80,000 Hours.

I think one consistent prevalent misunderstanding that people have of the 80,000 Hours advice is that they don't realize that 80,000 Hours' advice is on the margin. So they think that 80,000 Hours' ideal situation would be "Everyone in the world, ideally, follows our advice and goes into these careers." And then the audience hears that, or you know, imagines they're hearing that, and goes "Well, if everyone did that then all these bad things would happen. We wouldn't have people exploring new frontiers or doing exploratory research that doesn't have a specific goal, et cetera, et cetera."

But actually all along, 80,000 Hours has been like "No, on the margin, given the current allocation of resources of human capital around the world, here's what seems undervalued, and good."

And it's kind of similar to what you're saying the AI risk argument is: That, at the very least, we can say that on the margin it would be good to have more people working on this, or thinking about it.

Kelsey: Mm-hmm. Yeah, I think that's a big part of it, it's just that's not a very intuitive mode of thinking for people.

And it's hard when someone's making an argument to tell whether they're making an argument about the margin, or whether they're making an argument about the ideal distribution, or what.

Julia: Right, exactly. Are there any objections that you think are just based on a misunderstanding of the AI risk argument?

Kelsey: So I know some people seem to think that concern about AI like, originated with Eliezer Yudkowsky, and is pretty much exclusive to effective altruists who came at it through that route.

And they've found it really persuasive just to learn that lots of other people independently reached that conclusion. And that Stephen Hawking did not get it from Eliezer. And that Nick Bostrom seems to have, in parallel, reached many of the same conclusions. And that back when computing was just starting, Alan Turing and [I. J. Good] were all saying "Wow, this is where we're gonna go eventually, although who knows when."

And I think quite a few people I've talked to find it persuasive, that this was something lots of different intellectual currents of thought converged on. Because they've been under the impression that it was sort of this one weird quirk of the effective altruist community.

And if it were, then you know, even if you found the arguments persuasive, that would in fact be a pretty good reason to be skeptical of them. Because intellectual communities can absolutely spiral around wrong ideas that are reinforced by local social norms and stuff. And if we were the only people who found this convincing, that would in fact be a reason to be sort of [skeptical].

Julia: Right, yeah. That's a good point. In the long version of your article at least, you say “It's tempting to conclude that there's a pitched battle between AI risk skeptics and AI risk believers. In reality, they might not disagree as profoundly as you would think.” Can you elaborate on that?

Yeah. So you certainly get statements from Yann LeCun, from Andrew Ng, from lots of people on the skeptical side, that sound very dismissive. They're very, “This is science fiction, we don't need to think about this.” And I think that contributes to the impression that the field has some people who are like “doomsday,” and some people who are like “oh, just shut up.”

If you dig into it a little bit more, what Yann LeCun is saying is:

“I think that AGI is probably more than 100 years away. I think that most efforts now to make it safer will be unproductive. I don't object to some people trying to think about these principles, and trying to lay some groundwork in time -- but you know, since I believe this is hundreds of years away, that hardly seems like a good priority. And also some people are hyping this out of all proportion and promising that, like, in 2030 they're gonna end death and colonize the galaxy. And I wish they'd stop that.”

And I think there's really only a couple substantial disagreements between that position and the AI risk “very nervous” position. The AI risk “very nervous” person, I think would say:

“Yeah, I think it might be sooner than we expect. It might be hundreds of years away. But there's some reason to expect that actually it could happen to us a lot faster than that. And secondly, I think there's more potential for work now to matter than you seem to think.”

And you know, that disagreement is substantial. But it's a lot smaller than, you know, the accusations of science fiction nonsense, and the accusations of burying your head in the sand, necessarily get at. Like, people disagree on how much we can do now, and how far away it is. But almost everybody thinks that artificial general intelligence is possible, and almost everybody agrees that it will be dangerous and complicated. They just disagree very significantly on when it's going to happen -- and therefore on whether the stuff we're doing now could matter.

Julia: I would go even farther than that actually, and say that most people, even the people who are usually counted in the skeptical camp, agree that AGI is not just possible but likely to happen, likely to be developed at some point.

Kelsey: Yes. You're right about that. I think LeCun has said that he does think we will get AGI, just not for a while.

Julia: Yann LeCun?

Kelsey: Yes.

Julia: Why do you think there is so much apparent disagreement, given the, you know, more moderate amount of actual disagreement? Why do we keep getting in this situation where the AI risk “skeptics” and “believers” keep arguing past each other?

Kelsey: So I think a lot of that, is that much of this is happening in news articles that will try and get a skeptical quote, and not necessarily expand on all the depth there.

I think part of it is that a lot of people [generalize from], “I'm confident this will happen, but not for another century” to “This is science fiction nonsense.” They don't have a good way of evaluating that differently from, you know, “Maybe faster than light travel is possible.”

Centuries away are just very hard to think about. And then part of it is that there certainly is a lot of AI hype and nonsense out there. From… Every startup's claiming they're doing AI when they're doing linear regressions on their 200 data points. To like, yeah, some very bold claims. Which I'm actually hesitant to call “excessive hype” until they've failed to be borne out -- but certainly very bold claims, from Open AI, and Deep Mind, about what they're gonna be capable of within the next decade.

Julia: Right.

Kelsey: And I think those have made a lot of people sort of react against the hype by being like, “No, calm down, it's nonsense. AI can't do any of those things.”

Julia: Right.

Switching tracks a bit now, to another article that you wrote, along with Dylan Matthews, at the beginning of this year… it was titled “16 Big Predictions About 2019, From Trump's Impeachment to the Rise of AI.” And in this piece, you did what I wish journalists would do all the time, but in fact almost never do: you made falsifiable predictions, about important things, with probabilities attached.

Can you share an example of a prediction that you made? And some of your reasoning for how you picked that probability?

Kelsey: Yeah. So one that's been on my mind a little bit recently is I said I think there's an 80% chance that there won't be a recession this year.

I did some re-reading of Tetlock, freshening up to publish these predictions, just trying to remember what all the advice on doing it right is. This is very nerve racking, because it's our first time making predictions. Making predictions is hard. We will probably not have incredibly good calibration, we will certainly make some predictions that are false, because we made a lot of them. And even if we did have perfect calibration, some of them would be false.

Julia: Right. And with only 10 -- even if you are perfectly calibrated, there's still a pretty decent chance you'd look poorly calibrated on a sample size that's small.

Kelsey: Yeah, so I was very nervous about not looking even worse than, sort of inevitably well. And I do want to say, the prediction community was great. They embraced this. They said, "Hey, you can criticize these predictions now, but don't criticize them in December unless you're willing to criticize them now."

I definitely felt like they understood the concept that you’ve got to socially reward attempts at something if you wanted to have them. So that was cool.

Julia: That's great, good. I want to socially reward them for socially rewarding you.

Kelsey: Yeah, so good for them. So anyway, one of the pieces of advice was just to do more reference class forecasting than you'd naively feel comfortable with.

So instead of asking the question, "Is there going to be a recession?" by going like, "Well, there's a government shutdown, and it's been a while since the last recession, and I have a bad feeling about this year"... You go, "Okay, if I predicted a recession, in every year for the last couple decades, how often would I have been right?" Turns out you're right about 15% of the time.

I bumped it up from there to 20, because it has been a long time since the last recession and there were some economic indicators a little bit suggestive that things looked a little bit worse than maybe the baseline. But that had very little influence on the estimate compared to how much of it was just, "All right, well if I'd been making this prediction every year, how well would I have done?"

Julia: Right.

Kelsey: Which feels weird but that's sort of the recommended starting point if you want to make predictions.

Julia: Yeah, interesting. I was reading your interview on 80,000 hours a little while ago. And I sort of smiled at this part where you mentioned this 80% prediction, and you said you've been feeling nervous when it looked like there might be a recession, because you had put 80% probability on there not being one. And you were like, "Gee, it's a little disturbing to notice that the reason I'm rooting against a recession is because I don't want to be proven wrong, as opposed to all the human suffering it would cause!"

Kelsey: Yeah.

Julia: But I'm very sympathetic. This is one of the big reasons why no one wants to make forecasts -- because they're afraid they're going to look bad if they're proven wrong, and they're going to have to stress about it, and so on.

I was wondering if you have any tips for how to overcome that fear. You're socially rewarded by people, but if you can't count on that, or in addition to that, what would you suggest?

Kelsey: Yeah, it's super hard to be wrong. I think it's just something you have to do a lot of deliberate practice at. When you're wrong, going, "Well, I learned that I was wrong. And I'm glad of that, because it will let me do stuff better."

I do believe that having stuff you really want to accomplish makes it easier to be wrong, because it's easier to go, "Well, now I have the information I need to be right." Whereas if you're sort of doing this for pride or doing it for its own sake, then your pride is always going to take a hit when you're wrong.

Julia: Do you think it would be feasible to build predictions into articles or op ed pieces? Anytime you, or a freelancer writing for Future Perfect, makes an argument in their piece, could you have them make a corresponding falsifiable prediction or two, that’s sort of logically or evidentially related to the argument?

Because that to me is kind of the dream. Even beyond having people do an annual or monthly batch of forecasts.

Kelsey: Yeah, that would be amazing. And you could have by someone's byline, their calibration score, so everybody knows how seriously they would take it.

Julia: Yes, that's part of the dream.

Kelsey: Yeah, well, I think it would add some time to articles. At least at first, I don't think it would produce a huge uptick in any of the metrics that journalists are incentivized to care about, unfortunately.

Julia: Yeah.

Kelsey: I do think it would be valuable. I would be pretty excited about figuring out how to make it happen.

One point to make is that formulating a prediction precisely is really challenging. And often I think I have a pretty precise formulation. And then I run it by someone and they're like, "I don't know what you mean, what about these three cases that this fails to differentiate between?"

So it's hard, and maybe a whole separate skill of its own, to specify a prediction clearly enough that it gets out what you mean, and has a single interpretation that everybody's going to agree on.

Julia: Right. Yeah. All right, I want to make sure we have time to talk about some of the posts on your Tumblr that I particularly liked. So just to remind our listeners, that blog is theunitofcaringtumblr.com. That’s all one word, theunitofcaring.

And one thing that I like about your writing on your blog, Kelsey, is: I like how you really take seriously both ethical questions, but also questions about people's mental health and personal flourishing. I mean, it's pretty rare. Well, I guess it's pretty rare for anyone to really take seriously either of those, but it's especially rare for someone to take both seriously, and take seriously the potential tensions between those two things.

So along those lines, one post of yours that stuck with me was about why it's not necessarily always good to just read arguments you disagree with. There's this common wisdom about, “You should seek out and engage with arguments from people you strongly disagree with. That's how you grow and change your mind, the virtuous thing to do.”

I mean, many people feel almost morally obliged to do that. So what's your case against that?

Kelsey: Yeah, so I think I often see people reading someone they strongly disagree with, and it makes them less charitably inclined towards the ideas. It makes them more angry and more defensive, if you immerse yourself in it. It can give you this perception that these ideas that you hate and that are wrong are on the rise and going to destroy everything you love.

And I see this on all sides. I see people who are liberal engage with conservative sites and become really furious about how horrible conservatives are.

I see conservative sites that link harmless articles giving advice to trans teenagers, and then everybody gets outraged and horrified about them.

I see conservative Catholics reading sex advice guides just to be really miserable about the degeneracy in the world these days.

And I think these people are thinking, if you're listening to the other side, even if you end up disagreeing, then you've done something virtuous.

But this isn't virtuous. It's self destructive. I don't think it teaches you very much about people. And I certainly think that if they are right, you will never learn that they are right by doing this.

So yeah, the advice I gave instead was: Find somebody who you respect a lot and admire, and you feel like you have a lot of things to learn from them, who disagrees with you about something. And this will make you more charitable towards the idea where you two disagree. And they will probably be a good person for you to learn about that idea from,' because you have this baseline respect for them.

And that's how to expose yourself to ideas you disagree with, is through a speaker who you respect and who you think of as on your side in some important ways.

Julia: I strongly endorse this advice, and I've given it myself, possibly inspired by your posts. I honestly don't remember at this point how much of my ideas are my own and how much are inspired by people I've read, so apologies if I've stolen any of it.

But anyway, when I've made this point, sometimes I get the pushback that, "Well, doesn't that mean you could only ever change your mind a little bit?" Or you wouldn't change your mind about underlying premises, because you've selected for people who already agree with you about those things, because those are the people you respect, and so on.

Kelsey: I mean, I can respect people who I have some pretty profound disagreements with. I have a very good friend who's an effective altruist and she's Catholic. That's a really substantial disagreement. But since we're both really interested in making the world the best place we can be, and both donating a lot of money to effective charities as one route to accomplishing that, and both interested in achieving lots of the same goals in other domains -- this makes me more inclined to have an open mind and listen to her about Catholic perspectives on things. And I don't think the disagreement there is small.

Julia: Do you ever change your mind, or moderate or modulate your position, in response to the Catholic arguments? Because I guess that violates my model of how to change your mind, which is that, you should be seeking out not just people you respect or like on a personal level, but people who sort of share your core premises about how to think and what kind of evidence counts and so on.

And that if I'm talking to someone who's against abortion, and their arguments are religious or are truly deontological or something, and I'm a consequentialist, there's just not a lot for us to engage with with each other.

Kelsey: Yeah, I think lots of Catholic EAs are happy to discuss abortion in terms that makes sense to the consequentialists around them. And I've had lots of those conversations, and I do think they've made me more pro-life. Not in the sense that I think the US government should be throwing people in jail for having an abortion. But in the sense of me thinking it's more probable than I used to think that an abortion is a fairly bad outcome, which we should be motivated in policy to try to minimize.

Julia: Interesting. Because of -- is it easy to summarize?

Kelsey: Yeah, because of talking to people who have strong moral intuitions in that direction, and coming up with thought experiments that articulate their intuitions, and making comparisons to other kinds of minds that I value.

I think if I were to summarize the update, in our language, I would say: “I'm very uncertain right now about what kinds of minds have the property that when they die, it is bad. I think when humans die it is bad. I think when animals die, it is probably fine. But it wouldn't actually shock me, if I had full information about the experience of being an animal, if I was eventually like, ‘Oh, no, it's actually also bad when animals die.’

And similarly, it seems possible that if I had a full understanding of the experience of being a fetus, that I would end up going, ‘Oh, yeah, this is the kind of mind where something tragic has occurred when this dies.’”

Julia: And this is a different question, or separate question, from the suffering question, I presume? You do think it's bad when animals suffer. It's just the question of whether it's bad when they die, is a different question. Kind of hard to think about.

Kelsey: Exactly. Much harder to think about. With suffering, I can kind of go, "Okay, do they have the same neural structures for experiencing pain that I do? Okay, they probably experience pain, probably they experience pain in much the same ways that I experience pain."

And there's really no reason to think that the experience of being kicked in the ribs varies between a dog and a human. Given how much of the structures we have to experience it.

But that doesn't answer the question of whether it is a bad thing when a dog dies. And that's, first I think, a question I'm confused enough about that pro-life friends would sort of able to convince me, "Hey, you should be really confused about whether it's bad when a fetus dies." And I was like, "Yeah, all right. I'm convinced that I should be really confused about that."

Julia: Yeah, it's really interesting.

Another one of your posts that has stayed with me is a post in which you were responding to someone's question -- I think the question was, “What are your favorite virtues?” And you described three. They were compassion for yourself; creating conditions where you'll learn the truth; and sovereignty.

And I wanted to ask first about that second one, “creating conditions where you'll learn the truth.” It's an interesting phrasing, because it's kind of adjacent to but different from these two much more common ideas that are already in the discourse, of: one, seeking out truth, going out and investigating things… and two, being willing to change your mind, or update, when you're confronted with new evidence or argument.

So can you talk about why you specifically picked “creating conditions where you'll learn the truth” instead of seeking out truth, or being willing to change your mind?

Kelsey: Yeah, so part of that is that I think that being willing to change your mind and seeking out truth are both very hard virtues to practice. And virtues where it's kind of easy to deceive yourself as to how well you're doing at them. Because you can tell yourself that you're very willing to change your mind, and just haven't run across things worth changing your mind about. And you can change your mind about things that don't matter very much. While still having important parts of your worldview that you sort of aren't actually up for criticizing. And it's hard to tell from the inside whether you're doing that.

Whereas I think it's pretty easy to tell from the inside whether you're creating conditions under which you can learn the truth. You can ask yourself, "How many friendships do I have? How many blogs do I read?

How many books do I read? How many podcasts do I listen to where people say things I profoundly disagree with, that make me think?" You can ask yourself, "When I encounter a question that makes me wonder if I'm wrong, do I keep learning and keep thinking? Or do I stop there and say, 'Well, that's enough'?"

So, in some ways, I prefer it as a virtue just because I think it's more concrete to answer the question, "Am I practicing this virtue?"

Julia: Right. That is a good virtue of a virtue -- is it concrete to answer, "Am I practicing it?"

Kelsey: Yeah. I think if people want to become more virtuous, it's good to throw virtues at them that they can tell if they're doing it right or not.

Julia: Right. The other virtue I wanted to talk about was sovereignty. Because I bet it will be less ... It's just less discussed. But it seems really important. And I only have realized in the last few years how important it is and how many people lack it, in important domains.

Can you explain briefly what sovereignty means?

Kelsey: Yeah, so I characterize sovereignty as the virtue of believing yourself qualified to reason about your life, and to reason about the world, and to act based on your understanding of it.

And I think it is surprisingly common to feel fundamentally unqualified even to reason about what you like. What makes you happy. Which of several activities in front of you, you want to do. Which of your priorities are really important to you.

I think a lot of people feel the need to answer those questions by asking society what the objectively correct answer is, or trying to understand which answer won't get them in trouble. And so I think it's just really important to learn to answer those questions with what you actually want and what you actually care about.

Julia: Can you give an example of a situation in which someone might want to defer to what the "correct" answer is, to what they should want?

Kelsey: Yeah. Say somebody is contemplating whether to get married. It seems very common to think about, "Well, what will people think of me if I don't do this? What will it say about me if I'm unmarried at my age? What will it say about me if I get married at this age? How mad will people be at me if I do this?"

And it can be hard to sort of focus on, as your overriding consideration, "What do I want? What does my best life look like? And is this the path to it?"

That's one that this comes up in everybody's life. But I think similarly, in effective altruism, a lot of people try to figure out what they should be doing, try to figure out what following other people's advice looks like for them, and really struggle with going, "Okay, what outcomes do I want? What actions put me on a path there? And what do I actually believe I should be doing?"

And it seems to me like the same sort of mistake.

Julia: I shared this post of yours on Twitter a while ago, and specifically pointed to the sovereignty point. And Rob Wiblin objected that, well, maybe you shouldn't have sovereignty on questions where your own judgment's less reliable than the consensus of the relevant experts. What's your reaction to that?

Kelsey: So the problem is, I don't think you can have that shortcut, even if it would be nice. You still have to figure out who the relevant experts are. And you still have to figure out in which areas your judgment isn't that good?

I think it is important to have good societal defaults. I think it is important that if somebody is the kind of person to just defer to the consensus on every question, that we as a society have good enough consenses that this doesn't screw them over.

But fundamentally, as an individual thinking, you can't really do that. There's no consensus sitting around to be a reasonable backstop, and no reasonable way of telling when you should or shouldn't defer to it. You still have to do the work of saying, "Okay, I think I'm going to defer to experts." And I do defer to experts all the time.

I think my understanding of sovereignty is very compatible with saying, "On this question, I just completely trust this researcher, and whatever answer they come up with, I think they're probably right." But you have to decide why you trust that researcher in particular.

Julia: One insight that I had from reading your post in particular was that maybe a lot of debates over whether you should "trust your gut” are actually about sovereignty. I was always very dismissive when people would say things like, "Oh, you should trust your gut, trust your intuition." Because basically I was imagining someone trusting their intuition about vaccines causing autism, as opposed trusting the scientific evidence.

But now, I wonder whether maybe a lot of the term “trust your gut” just means, "Well, take your preferences into account because they are important data." Or pay attention to your hesitation around deferring to a particular expert, and actually try to figure out for yourself which experts are trustworthy, or something like that.

Kelsey: Yeah, I definitely think -- maybe replace “trust your gut” with --

Julia: Consult?

Kelsey: Yeah, check in with your gut. Treat your gut as some information.

Julia: Yeah.

Kelsey: And making your gut more informative is an important part of your growth as a person.

Julia: Right. That's actually very well put. Because I do trust my judgment quite a lot. I think I have sovereignty in a lot of domains, although not all domains. But I think one of the reasons I have that is that I've formed opinions, and then I've found out whether they were right or not, and I've revised my thinking, and over time, I've kind of developed some trust in my judgment -- but it wasn't trust by default.

Kelsey: Yeah, I think my process has been similar. I’ve stewed over lots of hard questions. And I got a sense of when I've tended to be right, and when I tended to be wrong, and that informs my gut and the extent to which I feel able to trust it now.

Julia: Right. Well, this was an unintentionally good segue for me into the last thing I wanted to talk about your Tumblr, a thing that I really like that you do often that many people don't do: you steelman arguments.

A thing you do sometimes is someone will submit a ... Is it an ask? I don't really know Tumblr, I'm just like a lurker who reads other people's Tumblrs. But there's this thing called an “ask” where people submit a question or prompt or something.

Kelsey: Yeah.

Julia: And then you answer it. So anyway, there will be an ask where someone has some exaggerated, straw-manny, inflammatory position that they want you to respond to. Like, “It's terrible that society is brainwashing kids into thinking they're the other gender and that they should chop off their genitals -- isn't this terrible? How can you not think this is terrible?” Or something.

And you'll respond to stuff like this in this very calm and measured way, like, "Okay, I'm going to pretend you didn't ask the question in that extremely unnecessarily inflammatory and kind of exaggerated way. Here's a concern that feels sort of what might be at the root of what you're talking about, that is a more reasonable concern someone might have. And here's why I still disagree with that."

And that just feels, it's (a), so much more interesting to read. And I can imagine that it would also be much more interesting and convincing to people, to readers who are maybe not the original submitter, but at least kind of on the fence or confused about the topic. It's more useful for them to hear an answer to a reasonable question than to hear the answer to the original unreasonable question, which would just be like, "Oh, my God, stop straw-manning."

And I've seen you talk also about how steelmanning is kind of a guiding principle of the work you do at Future Perfect.

But my questions for you is: When I talk to people about steelmanning, I sometimes get objections that it might actually be a bad thing. One of the objections is, "Well, isn't Steel-manning going to cause you to be overly charitable or sympathetic to views that are actually bad or dangerous?" That you'll just kind of assume people mean the more reasonable thing, but actually, they mean the unreasonable thing, and their unreasonable view is bad and dangerous and should be combated or stomped out.

And then the other objection is: Steelmanning might actually be bad for you, in that in the process of trying to find the more reasonable interpretation of what someone said, you might actually miss the point they're trying to make, because it's not the thing that most immediately seems reasonable to you.

Do either of those concerns seem reasonable to you?

Kelsey: Yeah, I think they do. I think it's sort of unfortunate that we have one word for both an internal technique for trying to understand perspectives you don't understand before, and an external rhetorical technique for trying to engage productively with a bad argument. And I think you need sort of different skills to employ each of them usefully.

As a rhetorical technique, I think the most important thing you need to be able to do is imagine you have an audience reading this post, and they flinch when they read the awkward inflammatory unreasonable framing, because they're like, "Oh, yeah, I sort of feel that way. But I wish people would ever say it outright who weren't jerks, who say it in this inflammatory way."

And I think, if you have a good sense of your audience, and you have an accurate, well-calibrated sense of who's flinching and what they believe and which had been articulated, then you can be highly effective by articulating it for them, and saying, "Yeah, what if we were talking about this? Because we should talk about this? Sure."

So that, the sort of way it goes wrong is if you don't understand your audience, and you don't actually get what people are hoping will be said. Which is probably a mistake I make sometimes, and it's certainly a mistake I witness a lot, is somebody assumes that the steelman... is this argument that they think will be very compelling to most of their readership. And then actually, the people who kind of hold that perspective are like, "Wait a second, that's not that's not a steelman. That's just a different argument."

Julia: Right, yeah.

Kelsey: And so that's the way that one fails. And then the way the internal one fails, I guess maybe it's kind of similar. But I think if the internal one is failing, if you instead of understanding the thing that they're trying to say, you just come up with something that's [reasonable] enough in your own worldview, but it's actually missing critical components of what makes it work as an argument.

And then you're like, "Well, this is a bad argument." And then you feel free to dismiss it, because the strongest version of it was bad -- when in reality, it was more that it was integrated into a different worldview. And when you chopped it out of that worldview, and brought it into yours, then it didn't have anything holding it up anymore.

Julia: Right. Well put. I like that metaphor. I'm imagining something being transplanted out of its native climate, and withering.

Kelsey: Yeah, exactly.

Julia: So we're almost out of time. Before we wrap up, Kelsey, I wanted to ask you for a recommendation, or just a nomination of a book, blog article or other resource, or even a thinker, a person who you have substantial disagreements with, but nevertheless have gotten value out of reading or engaging with?

Kelsey: Yeah. So one blog I've gotten a ton of value out of engaging with recently is Andrew Gelman's blog on statistical significance and methodology in the sciences.

And the main thing I get out of it is that he'll post lots of papers and break down their methodology. And he comes down on the, “we should just abolish statistical significance" side of things. I don't think I do. But I have picked up so many mental tools from just reading through what he's doing. And now when I read a paper, I think I have a little shoulder Andrew Gelman who's like, "That effect size looks suspicious. That seems like you've got lots of comparisons probably went into that set of results you just reported there. This looks fishy." And I think everybody should have one of those. If you're going to be reading any papers.

Julia: A little shoulder Andrew Gelman?

Kelsey: Yes, definitely.

Julia: Nice.

Kelsey: So I highly recommend his blog, to pick up one of those.

Julia: And just to remind our listeners, you can read Kelsey's work at Future Perfect, as well as Dylan's work and the work of the other freelancers who contribute. There's also the Future Perfect podcast that you should check out. And we'll link to some of Kelsey's articles and blog posts that we mentioned during the episode and we'll link to Andrew Gelman's blog as well. Great. All right. Well, Kelsey, thank you so much for coming on the show. It's been such a pleasure having you.

Kelsey: Yeah, thank you so much.

Julia: This concludes another episode of Rationally Speaking. Join us next time for more explorations on the borderlands between reason and nonsense.