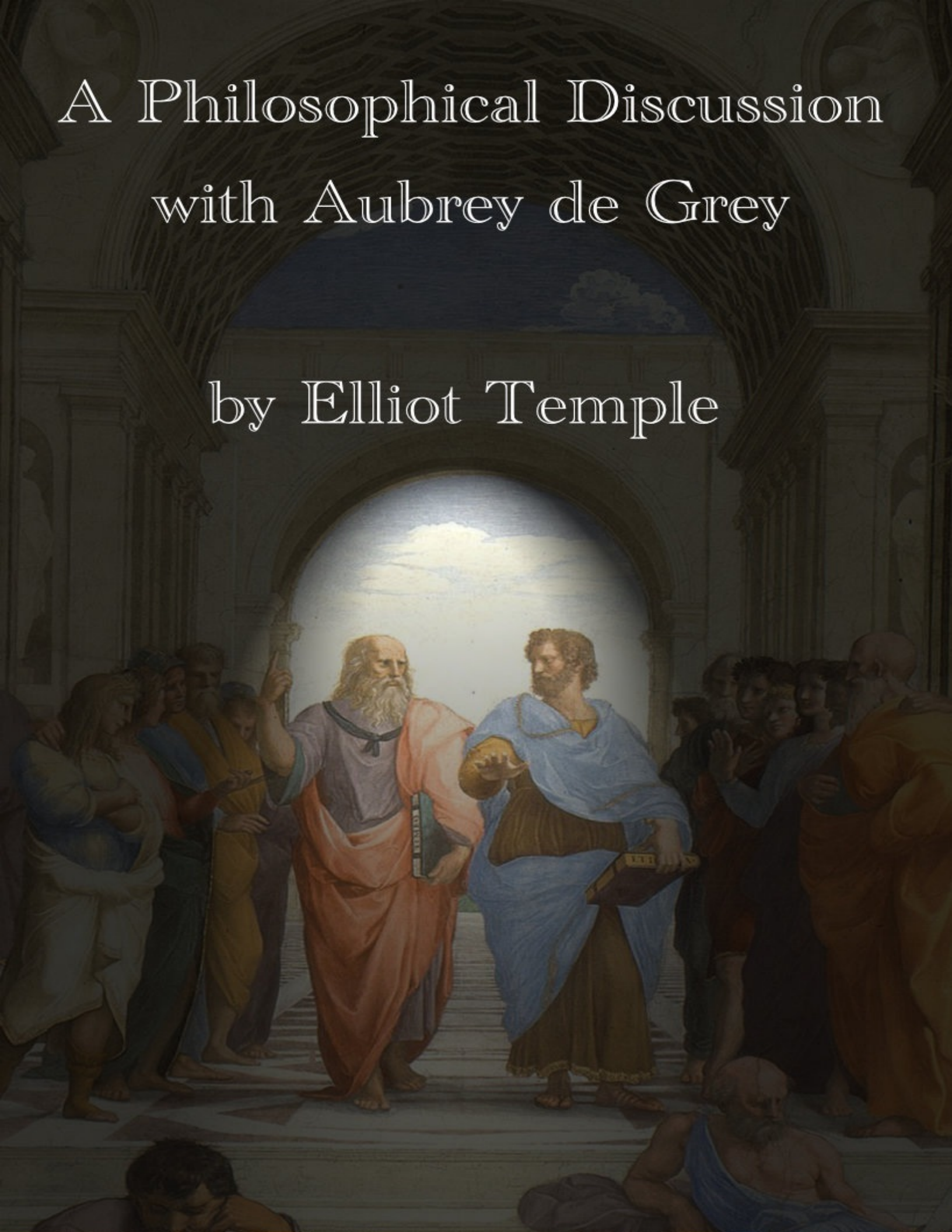
A Philosophical Discussion with Aubrey de Grey

by Elliot Temple



A Philosophical Discussion with Aubrey de Grey

By Elliot Temple

elliot@fallibleideas.com

www.fallibleideas.com

www.curi.us

Editor's Introduction

This book collects a wide-ranging email conversation between philosopher Elliot Temple and scientist Aubrey de Grey. Topics include knowledge, cryonics, and philosophy.

Elliot Temple is an American philosopher whose intellectual influences include David Deutsch, Ayn Rand, Karl Popper, William Godwin, and Ludwig von Mises. He has made several important original contributions to philosophy, including the idea of **Paths Forward** and **Yes or No Philosophy**.

Aubrey de Grey is a biogerontologist and the driving force behind **SENS** – Strategies for Engineered Negligible Senescence. SENS is an organized and comprehensive medical research program to deal with the problems caused by aging. If you're interested in SENS, read Aubrey de Grey's book **Ending Aging**.

Quote Coloring

- Yellow-highlighted quotes are from Aubrey de Grey.
- The text with no colored highlights is Elliot Temple talking presently.
- Bluegreen-highlighted quotes are from Elliot Temple at a previous point in the discussion.
- Red highlights are quotes from other sources (such as websites linked to in the discussion).

-Justin Mallone

Aubrey de Grey Discussion, 1

I began the discussion like this:

You endorse Alcor and CI:

http://www.reddit.com/r/Futurology/comments/28e4v3/aubrey_de_grey_ama context=5

> For the millionth time let me stress that referring to "getting older without getting sicker" as "becoming immortal" is not only inaccurate but actively counterproductive to this mission, because it entrenches the view of skeptics that the mission is quixotic. To answer the question you should have asked: obviously it depends on your age, but absolutely, everyone should have a life insurance policy with Alcor or Cryonics Institute, for exactly the same reason that they should have any other kind of health insurance.

Take a close look at Alcor and CI. While cryonics is a good idea in principle, Alcor and CI have lots of big problems (including that current cryonics technology isn't really good enough).

One big problem is not freezing people quickly. Max More, President and CEO of Alcor, writes:

http://lesswrong.com/lw/bk6/alcor_vs_cryonics_institute/69z7

> You mention Mike Darwin, yet note that in Figure 11 of a recent analysis by him, he says that 48 percent of patients in Alcor's present population experienced "minimal ischemia." Of CI, Mike writes, "While this number is discouraging, it is spectacular when compared to the Cryonics Institute, where it is somewhere in the low single digits."

Alcor CEO brings up, favorably, a statistic meaning that Alcor does a bad job at least 52% of the time. Because, hey, CI does much worse, and the discussion

topic is a comparison.

So I don't think you should tell people to sign up for CI and suggest it's the same quality as regular medicine.

You can find lots more information:

http://lesswrong.com/lw/bk6/alcor_vs_cryonics_institute/

http://lesswrong.com/lw/343/suspended_animation_inc_accused_of_incompe

(Comments include discussion from people like former Alcor President Mike Darwin.)

http://www.alcor.org/cases.html

http://www.cryonics.org/case-reports/

See e.g. the most recent CI case:

http://www.cryonics.org/case-reports/the-cryonics-institutes-123rd-patient

> CI patient #123 was a 71 year old male from England. Due to the uncontrollable circumstances of this case, the patient was straight frozen without being perfused with cryoprotective solutions and was sent to the Cryonics Institute for long-term storage in liquid nitrogen.

They failed. As they often do. No cryoprotectants! And they don't care to provide details. And they indicate they won't do anything different in the future, since they consider whatever happened "uncontrollable".

The latest Alcor case is very problematic too:

http://www.alcor.org/Library/html/casesummary2680.html

They argued with a Medical Examiner for a while, then managed to get ahold of the body and began cool down 2.5 days after death. The delay sounds very worrisome to me, but the case report doesn't address this problem at all. No medical details are provided about how cool down went. And there's no

explanation about what temperature the body was at for the 2.5 day delay, the resulting damage, and whether this person could reasonably be expected to ever be revived.

I like SENS. I like life. I like the idea of cryonics. But I wouldn't pay a bunch of money for the bad patient outcomes which CI and Alcor routinely provide (according even to their own claims on their websites).

Aubrey de Grey Discussion, 2

> I don’t understand your logic here. I’m well aware of the issues you mention regarding the quality of Alcor’s and CI’s preservations, and I’ve never suggested that any current cryonics service is the same quality as regular medicine. Why do you think it would need to be that good to justify signing up?

I don't think it would have to equal regular medicine to be worthwhile. But the gap is big, and cryonics is expensive.

You said everyone should sign up for cryonics, for the same reason they have regular health insurance. This suggests that cryonics has traits seen with regular medicine, like being run pretty competently, providing value for cost, routinely providing good outcomes, and making your life better. Cryonics currently provides none of those.

To answer your question about what would justify signing up: First, I'd want cryonics organizations to be run in a competent and responsible way. Second, I'd want cryonics technology to improve enough to preserve brains well enough to optimistically expect the relevant information (about one's mind and ideas) to be preserved, and I would want cryonics organizations to provide quality persuasive intellectual explanations on this point. I think those two problems are deal breakers.

Regarding preservation, without staff errors, one big problem is fracturing – meaning breaks in the brain. Alcor's attitude seems to be that fracturing doesn't destroy information and nanotech can theoretically fix it because the breaks are smooth and the separated parts of the brain do not end up far apart. I'm not convinced; I think they'd need much better reasons to say this physical brain damage is OK and the relevant information still preserved. (I also think the idea of nanotech repairs is misguided. The focus should be on one day getting the information from the brain into a computer, not on fixing and reviving the original organic brain.) Fracturing is not the only serious technological problem.

If those two issues were fixed, I still would not recommend cryonics to "everyone", or most people, because it'd be a large financial burden for most people on Earth, in return for a long shot. Unless cryonics improved SPECTACULARLY, it wouldn't be worth signing up at a big cost to one's standard of living now. There's also the issue that the majority of people don't value life and don't want to live, in some pretty fundamental philosophical ways, as explained e.g. in Atlas Shrugged. Cryonics, like SENS, doesn't fit everyone's values and preferences.

It would also help if societal institutions handled cryonics better, e.g. if you could conveniently go [to] a cryonics facility and kill yourself on site with staff present, rather than having them wait around for you to die (possibly suffering increasing brain damage from your disease in the meantime), wait for you to be pronounced legally dead, and perhaps deal with days of interference from regular medical personnel. Similarly, sometimes courts order people removed from cryo facilities. These things lower the chance of getting a good patient outcome, but I don't see fixing this as a strict requirement to sign up.

It would also be nice if I was a lot more convinced that Alcor and CI won't go out of business within the next 50 years, let alone 1000 years. Cryo preservation requires frequent maintenance and upkeep costs.

> Two more points:
>
> - A key feature that you don’t mention is that the poor preservations you list are cases where the individual did not do what I also strongly recommend, namely get themselves to the vicinity of their provider while their heart is still beating. Other cryonicists’ self-neglect isn’t a very good basis for one’s own decisions.

I don't think you read the cases closely. The Alcor case said he was in the Phoenix area, which is around 12 miles from Scottsdale, where Alcor is. It is the vicinity. Alcor refers to the "Scottsdale/Phoenix metropolitan area" on their website when explaining why they chose their location.

The reason for that bad outcome, and bad case report writing, was not due to location. For the CI case, it doesn't say what the reason for the bad outcome was, so we don't know if it had to do with location or not.

There are plenty of cases where people did everything right and got bad outcomes. There are even plenty of cases where cryo personnel irresponsibly caused bad outcomes. I include an example at the bottom of this email. There are, unfortunately, more examples available at the links I provided.

> - As you say, current cryonics technology has a ways to go; but that's another reason to sign up, since the more members Alcor and CI have, the more they can work to improve the technology.

Signing up for medical purposes, and for donation purposes, are different.

You said that, "... everyone should have a life insurance policy with Alcor or Cryonics Institute, for exactly the same reason that they should have any other kind of health insurance."

Signing up because you want to donate is not signing up for "exactly the same reason" as one has regular health insurance.

And I do not think everyone is in a financial position where they should donate money to cryonics research (or to anything).

For a younger American signing up for Alcor, the rough ballpark cost is 35 minutes of minimum wage work, 365 days a year. That's a big deal. That is a lot of one's life! Cost increases with age, so that's a minimum. (CI costs less than half that, which is still a lot of money for most people, and the quality drops along with the price.)

And I think if people have the means to make medical donations, SENS is a better option than cryonics. The SENS project you explain very well in Ending Aging, and elsewhere, makes a lot of sense and is a great idea, and you're working on it in a reasonable, competent, and effective way. Cryonics is an in-principle good idea, but unfortunately it doesn't go much further than that today. And I don't think throwing money at the issue will fix problems like some of the bad ideas of the people involved with Alcor and CI.

Example of what can happen with cryonics, not the patient's fault:

http://www.cryonics.org/case-reports/the-cryonics-institutes-95th-patient

> Curtis deanimated under as favorable a set of circumstances as any of us could have hoped-for.

> A number of CI Directors have become concerned that I have been modifying the cryoprotectant carrier solutions without adequate testing ... In response to concerns by CI Directors (and my own concerns) I will not make more modifications to the carrier solutions, and I believe we should return to using the traditional VM−1 carrier for the time being

Ben Best, CI president (at that time), was experimenting on people who paid to be preserved. The result was failure to perfuse with cryoprotectants. And this is written by the guilty party. For an outside perspective, Mike Darwin comments:

https://web.archive.org/web/20120406161301/http://chronopause.com/index.p personal-identity-survive-cryopreservation/

> Even in cases that CI perfuses, things go horribly wrong – often – and usually for to me bizarre and unfathomable (and careless) reasons. My dear friend and mentor Curtis Henderson was little more than straight frozen because CI President Ben Best had this idea that adding polyethylene glycol to the CPA solution would inhibit edema. Now the thing is, Ben had been told by his own researchers that PEG was incompatible with DMSO containing solutions, and resulted in gel formation. Nevertheless, he decided he would try this out on Curtis Henderson. He did NOT do any bench experiments, or do test mixes of solutions, let alone any animal studies to validate that this approach would in fact help reduce edema (it doesn't). Instead, he prepared a batch of this untested mixture, and AFTER it gelled, he tried to perfuse Curtis with it. ... Needless to say, as soon as he tried to perfuse this goop, perfusion came to a screeching halt. [In other CI cases,] They have pumped air into patient's circulatory systems…

Ben Best and Mike Darwin discuss the matter further here:

http://lesswrong.com/lw/bk6/alcor_vs_cryonics_institute/6c35

Aubrey de Grey Discussion, 3

> I merely claim that even today we are good enough at it that those who help the providers to help them have a good enough chance of revival that it makes sense to sign up, even if the cost compares with that of traditional health insurance.

Can you point me to writing which you think makes a correct, reasonably complete (across multiple sources is fine), and persuasive case for this reasonable chance of revival?

If I'm mistaken about this I'd like to find out (and sign up for cryonics), and I am willing put in the effort to find out.

I don't agree it's a matter of "personal evaluation". There's an objective, impersonal truth of the matter about the current state of cryonics. Just like whether SENS is currently a good idea is a matter of objective truth, not of personal evaluation. And various people who disagree with SENS are wrong.

I think people should only sign up for cryonics if adequate, objective, pro-cryonics arguments/explanations exist, which they can read and see why it makes sense, and which include answers to all important criticisms. And if that does exist, then it'd be a mistake to disagree anyway as some kind of personal matter. I (like Popper, Deutsch and Rand, who have explained some of the reasons) don't go for that "agree to disagree" and "personal evaluation" type stuff, which can be a way to dodge the rational pursuit of truth.

> Let me conclude, however, by thanking you for your support of SENS and agreeing with you that SENS is plan A! It’s no accident that I work on SENS rather than on cryonics.
>
> Cheers, Aubrey

Yeah. Best wishes.

Aubrey de Grey Discussion, 4

> I can’t point you to anything better than what is posted at Alcor’s and CI’s sites, no. Instead let’s look at what you say below. Sure there is an objective, impersonal truth of the matter about the current state of any particular technology. The question is whether what we do with that truth should be similarly objective and impersonal, and I don’t think it should. I believe it is OK for people to have different values and priorities, whether it’s concerning the merits of tomato ketchup or the value of life. Therefore, I believe there is a range of legitimate opinions about the justifiability of a given course of action. For sure that range will be finite, i.e. there will be cases where people are not adopting a policy that is consistent with their other beliefs and will be resistant to recognising that fact, but that doesn’t change that fact that there is still that (finite) room for legitimate agreement to disagree. Cryonics is a rather extreme case, because its basis in the prospect of revival in the rather distant future entails so much uncertainty as to the pros and cons. I value my and others’ lives very highly, and I consider it quite likely that the future will be a progressively more fulfilling place to be, so I think signing up for cryopreservation makes sense even if one evaluates the chance of being revived and being glad one had been is quite low (I would probably go as low as 1%, and I definitely think we’re up at at least 10% today, even taking into account the issues we’ve been discussing). But I don’t claim to have an objective, impersonal argument for that 1% - rather, if someone else values life less than I do and/or they are more pessimistic about human progress, and they conclude that their cutoff is 50%, they’re welcome to their opinion. No?

I agree about some scope for people to differ, though I don't think the reasonable range extends to not signing up for cryonics that is 50% likely to work, for people who can afford it.

I, too, value life very highly and expect the future to be dramatically better. I think concerns about e.g. overpopulation and running out of jobs are bad philosophy, both generally (problems are soluble, and we don't have to and shouldn't expect to know all future solutions today) and also I could give specific

arguments on those two issues today. And I'm not worried that I might not be glad to be revived.

But we have a disagreement about methodology and epistemology, which comes up with your comments on percentages.

If I believed cryonics had even 1% odds in a meaningful sense, I'd sign up too. I value my life more than 100x the price. That's easy. An example of meaningful odds would be that for every 1000 people who sign up, 10 will be revived. But it doesn't work that way.

Explanations don't have percentage odds. It's important to look at issues in terms of good and bad explanations, and criticisms, not odds. (You may have some familiarity with this view from David Deutsch, including his criticisms of weighing ideas and of Bayesian epistemology.)

In FoR, DD uses the example idea that eating grass will cure a cold. Because there's no explanation of how grass does that, he explains that this empirically-scientifically testable idea isn't worth testing. It should be rejected just from the philosophical criticism that it lacks a good explanation.

It shouldn't be assigned a probability either. It's bad thinking, to be rejected as such, unless and until a new idea changes things.

Odds apply to issues like physical events. Odds are a reasonable way to think about the possibility of dying in a plane crash, or other cryo-incompatible deaths. Odds can even somewhat model problems like whether the cryo staff will make a mistake, or whether Alcor stays in business, though there are some problems there.

You could die in a plane crash, or not. It could go either way, so odds make some sense. But either current cryo methods (assume perfusion etc go well) preserve the necessary information, or they don't. That can't go either way, there's a fact of reality one way or the other.

The basic way odds are misused is there are multiple rival ideas, and rationally resolving the conflicts between them turns out to be difficult. So people seek ways to retreat from critical discussion and find a shortcut to a conclusion. E.g. a person favors an idea, and there is some idea which contradicts it which he can't objectively refute. Rather than say "I don't know", or figure out how to know, he

assigns some odds to his idea, then lowers the odds for each criticism he doesn't refute. But the odds are completely arbitrary numbers and have no bearing on which ideas are correct.

Fundamentally, he's mistaken to take sides when two ideas contradict and he can't refute either one. Often this is done by bias, e.g. favoring the idea he thought of himself, or spent the last five years working on.

A starting point for a cryo explanation is that digging up graves to revive people won't work, due to brain damage (this could be explained in more detail I won't go into). There is no good explanation of how it could ever work. This bad explanation isn't worth scientific testing, and should not be assigned any odds.

Freezing people is better than coffins because it preserves more brain matter and prevents a lot of decay, but there's no good explanation that it would work either, because there's so much brain damage. All claims that it would work can be refuted by criticism (in the context of present knowledge). But vice versa doesn't apply: one could write an explanation of why straight freezing won't work for cryo, which would stand up to criticism. (Today. All these things are always tentative, and can be rethought if someone has a new idea.)

That is how issues should be resolved rationally. Get a situation with one explanation that survives criticism, and no rivals that do. Then, while one could still be mistaken, there is a non-arbitrary opportunity to accept the best current knowledge.

This is a Popperian view, which many people disagree with. They're wrong. And all of their arguments have known answers. I can answer any points you're interested in.

Changing subjects briefly, let's apply this to SENS. SENS is the best available knowledge on the issues it addresses, and which should not be dismissed by arbitrarily assigning it odds. Odds are a semi-OK approximation for whether specific already-understood SENS milestones will be done by a particular date, but are not an OK way to judge the truth of the core explanatory ideas of SENS. It's very important to look at SENS in terms of the proposed explanations and criticisms, and actually resolve the conflicts between different ideas (e.g. go through the criticisms of SENS and figure out concretely why each criticism is wrong, rather than be unable to objectively and persuasively answer some

criticism but continue anyway. Note you are able to address EVERY criticism, which makes SENS good, as opposed to other ideas which don't live up to that important standard.)

Finally, today's vitrification processes cause less brain damage than freezing. But still lots of brain damage. So for the same main reason as before (lots of brain damage prevents reviving), cryonics won't work (until there's better technology).

Either this is the best available explanation, or there is information somewhere refuting it, or there is a rival for the best explanation that's also unrefuted. In each case, it's not a matter of odds, and this initial skeptical explanation regarding cryo I've given should stand as the best view on the matter unless there are certain kinds of specific relevant ideas (rivals, criticisms).

Behinds statements about odds, there usually are some explanations, but it'd be better to critically discuss them directly.

I'm guessing you may have in mind an explanation something like, "We don't know how much brain damage is too much, and can model this uncertainty with odds." But someone could say the same thing to defend straight freezing or coffins, as methods for later revival, so that can't be a good argument by itself.

To make a rational case for today's cryonics, there has to be some explanation about how much brain damage is too much, why that much, and how vitrification gets over the line (while, presumably, freezing and grave digging don't – though Alcor and CI don't seem to take that seriously, e.g. Alcor has dug up a corpse from a grave and stored it). Well, either there should be an explanation like I said above, or one explaining why that's the wrong way to look at it, and explaining something even better. Without good explanation, it's the grass cure for the cold again. You may also have in mind some further answers to these issues, but I can't guess them, and if they are good points that good content was omitted from the statement of odds.

Finally to put it another way: I don't think people should donate to SENS if the explanations in Ending Aging didn't exist (or equivalent prior material). Those good ideas make all the difference. Without those ideas, a claim that SENS might work (even with only 10% odds) would not suffice. And I don't think cryonics has the equivalent good explanations like SENS. (Though I'd be happy to be corrected if it does have that somewhere.)

If you are interested, I will write more explaining the philosophy here. Actually I did write more and deleted it, to keep things briefer. Epistemology, btw, is my chosen specialty. (I don't want any authority, I just think it's relevant to mention.)

Aubrey de Grey Discussion, 5

> I've been completely unable to get my head around what [David Deutsch] says about explanations, and you've reawakened my confusion.
>
> Essentially, I think I agree that there are no probabilities in the past, which I think is your epistemological point, but I don't see how that matters in practice - in other words, how we can go wrong by treating levels of confidence as if they were probabilities.

That thing about the past isn't my point. My point is there are probabilities of events (in physics), but there are no probabilities that ideas are true (in epistemology). E.g. there is a probability a dice roll comes up 4, but there isn't a probability that the Many-Worlds Interpretation in physics is true – we either do or don't live in a multiverse.

So a reference to "probability" in epistemology is actually a metaphor for something else, such as my confidence level that the Many-Worlds Interpretation is true. This kind of metaphorical communication has caused confusion, but isn't a fundamental problem. It can be understood.

The bigger problem is that using confidence levels is also a mistake.

Below I write brief replies, then discuss epistemology fundamentals after.

> The ultimate purpose of any analysis of this kind - whether phrased in terms of probabilities, parsimony of hypotheses, quality of explanations, whatever - is surely to determine what one should actually do in the face of incomplete information.

I agree with decision making as a goal, including decisions about mental actions (e.g. deciding what to think about a topic).

> So, when you say this:

> I'm guessing you may have in mind an explanation something like, "We don't know how much brain damage is too much, and can model this uncertainty with odds." But someone could say the same thing to defend straight freezing or coffins, as methods for later revival, so that can't be a good argument by itself.

> I don’t get it. The amount of damage is less for vitrification than for freezing and less for freezing than for burial. So, the prospect of revival by a given method is less plausible (why not less “probable”?) for burial than freexing than vitrification.

I explain more about my intended point here at footnote [1] below.

I agree that changing "probable" to "plausible" doesn't change much. My position is a different epistemology, not a terminology adjustment.

> But, when we look at a specific case (e.g. reviving a vitrified person by melting, or a frozen person by uploading), we need to look at all the evidence that we may think bears on it - the damage caused by fracturing, for example, and on the other side the lack of symptoms exhibited by people whose brain has been electrically inactive for over an hour due to low temperature. Since we know we’re working in the context of incomplete information, and since we need to make a decision, our only recourse is to an evaluation of the quality of the explanations (as you would say it - I rather prefer parsimony of hypotheses but I think that’s pretty nearly the same thing).

I actually wouldn't say that.

My approach is to evaluate explanations (or more generally ideas) as non-refuted or refuted. One or the other. This is a boolean (two-valued) evaluation, not a quantity on a continuum. Examples of continuums would be amount of quality, amount of parsimony, confidence level, or probability.

These boolean evaluations, while absolute (or "black and white") in one sense, are tentative and open to revision.

In short: either there is (currently known) a criticism of an idea, or there isn't. This categorizes ideas as refuted or not.

Criticisms are explanations of flaws ideas have – explanations of why the idea is wrong and not true. (The truth is flawless.)

Issues like confidence level aren't relevant. If you can't refute (explain a problem with) either of two conflicting ideas, why would you be more confident about one than the other?

When dealing with a problem, the goal is to get exactly one non-refuted idea about what to do. Then it's clear how to act. Act on the idea with no known flaws (criticisms) or alternatives.

Since this idea has no rivals, amount of confidence in it is irrelevant. There's nothing else to act on.

There are complications. One is that criticisms can be criticized, and ideas are only refuted by criticisms which are, themselves, non-refuted. Another is how to deal with the cases of having multiple or zero non-refuted ideas. Another is that parsimony or anything else is relevant again if you figure out how to use it in a criticism in order to refute something in a boolean way.

> And the thing is, you haven’t proposed a way to rank that quality precisely, and I don’t think there is one. I think it is fine to assign probabilities, because that’s a reflection of our humility as regards the fidelity with which we can rank one explanation as better than another.

I think there's no way to rank this, precisely or non-precisely. Non-refuted or refuted is not a ranking system.

I don't think rankings work in epistemology. The kind of rankings you're talking about would use a continuum, not a boolean approach.

I provide an explanation about rankings at footnote [2], with cryonics examples.

The fundamental problem in epistemology is: ideas conflict with each other. How should people resolve these conflicts? How should people differentiate and

choose between ideas?

One answer would be: whenever two ideas conflict, at least one of them is false. So resolve conflicts by rejecting all false ideas. But humans are fallible and have incomplete information. We don't have direct access to the truth. So we can't solve epistemology this way.

The standard answer today, accepted by approximately everyone, is so popular it doesn't even have a name. People think of it as epistemology, rather than as a particular school of epistemology. It involves things like confidence levels, parsimony, or other ranking on continuums. I call it "justificationism", because Popper did, and because of the mistaken but widespread idea that "knowledge is justified, true belief".

Non-justificationist epistemology involves differentiating ideas with criticism (a type of explanation) and choosing non-refuted ideas over refuted ideas. Conflicts are resolved by creating new ideas which are win/win from the perspectives of all sides in the conflict.

Standard "Justificationism" Epistemology

This approach involves choosing some criteria for amount of goodness (on a continuum) of ideas. Then resolving conflicts by favoring ideas with more goodness (a.k.a. justification).

Example criteria of idea goodness: reasonableness, logicalness, how much sense an idea makes, Occam's Razor, parsimony, amount and quality of supporting evidence, amount and quality of supporting arguments, amount and quality of experts who agree, degree of adherence to scientific method, how well it fits with the Bible.

The better an idea does on whichever criteria a particular person accepts, the higher goodness he scores (a.k.a. ranks) that idea as having. If he's a fallibilist, this scoring is his best but fallible judgment using what he knows today; it can be revised in the future.

There are also infallibilists who think some arbitrary quantity of goodness (justification) irreversibly changes an idea from non-good (non-justified) to good (justified). In other words, once you prove something, it's proven, the end. Then

they say it's impossible for it to ever be refuted. Then when it's refuted, they make excuses about how it was never really proven in the first place, but their other ideas still really are proven. I won't talk about infallibilism further.

This goodness scoring is discussed in many ways like: justification, probability, confidence, plausibility, status, authority, support, verification, confirmation, proof, rationality and weight of the evidence.

Individual justificationists vary in which of these they see as good. Some reject the words "authority" or even "justification".

So both the criteria of goodness, and what they think goodness is, vary (which is why I use the very generic term "goodness"). And justificationists can be fallibilists or infallibilists. They can also be inductivists, or not and empiricists or not. Like they could think inductive support should raise our opinion of how good (justified) ideas are, but alternatively they could think induction is a myth and only other methods work.

So what's the same about all justificationists? What are the common points?

Justificationists, in some way, try to score how good ideas are. That is their method of differentiating ideas and choosing between ideas.

One more variation: justifications don't all use numerical scores. Some prefer to say e.g. "pretty confident" instead of "60% confident", perhaps because they think 60% is an arbitrary number. If someone thought the 60% was literal and exact, that'd be a mistake. But if it's understood to be approximate, then using an approximate number makes no fundamental difference over an approximate phrase. Using a number can be a different way to communicate "pretty confident".

Popper refuted justificationism. This has been mostly misunderstood or ignored. And even most Popperians don't understand it very well. It's a big topic. I'll briefly indicate why justificationism is a mistake, and can explain more if you ask.

Justificationism is a mistake because it fundamentally does not solve the epistemology problem of conflicts between ideas. If two ideas conflict, and one is assigned a higher score, they still conflict.

Other Justificationism Problems

Justificationism is anti-critical because instead of answering a criticism, a justificationist can too easily say, "OK, good point. I've lowered my goodness (justification) score for this idea. But it had a lead. It's still winning." (People actually say it less clearly.) In this way, many criticisms aren't taken seriously enough. A justificationist may have no counter-argument, but still not change his mind.

Justificationism is anti-explanatory, because scores aren't explanations.

Another issue is combining scores from multiple factors (such as parsimony and scientific evidence. Or evidence from two different kinds of experiments) to reach a single final overall score. This doesn't work. A lot about why it doesn't work is explained here: **http://www.newyorker.com/magazine/2011/02/14/the-order-of-things**

One might try using only one criterion to avoid combining scores. But that's too limited. And then you have to ignore criticism. For example, if the one single criterion is parsimony, the score can't be changed just because someone points out a logical contradiction, since that isn't a parsimony issue. This single criterion approach isn't popular.

There's more problems, I just wanted to indicate a couple.

Popper Misunderstandings

A common misunderstanding is that Popper was proposing new criteria for goodness (justification) such as (amount of) testability, severity of tests passed, how well an idea stands up to criticism, (amount of) corroboration, and (amount of) explanatory power. This is then dismissed as not making a big difference over the older criteria. DD's (David Deutsch's) "hard to vary" can also be misinterpreted as a criterion of goodness (justification).

That's not what Popper was proposing.

Another misunderstanding is that Popper proposed replacing positive justifying criteria with a negative approach. In this view, instead of figuring out which

ideas are good by justifying, we figure out which ideas are bad by criticizing (anti-justifying).

This would not be a breakthrough. Some justificationists already viewed justification scores as going both up and down. There can be criteria for badness in addition to goodness. And it makes more sense to have both types of criteria than to choose one exclusively.

This wasn't Popper's point either.

Non-Justificationist Epistemology

This is very hard to explain.

Fundamentally, the way to (re)solve a conflict between ideas is to explain a (win/win) (re)solution.

This may sound vacuous or trivial. But it isn't what justificationism tries to do.

It's similar to **BoI**'s point that what you need to solve a problem is knowledge of how to solve it.

How are (re)solutions found? There's many ways to approach this which look very different but end up equivalent. I'm going to focus on an arbitration model.

Think of yourself as the arbiter, and the conflicting ideas as the different sides in the arbitration. Your goal is not to pick a winner. That's what justificationism does. Your goal as arbiter, instead, is to resolve the conflict – help the sides figure out a win/win outcome.

This arbitration can involve any number of sides. Let's focus on two for simplicity.

Both sides in the conflict want some things. Try to figure out a new idea so that they both get what they want. E.g. take one side's idea and modify it according to some concerns of the other side. If you can do this so everyone is happy, you have a non-refuted idea and you're done.

This can be hard. But there are techniques which make solutions always possible using bounded resources.

DD would call this arbitration "common preference finding", and has written a lot about it in the context of his Taking Children Seriously. He's long said and argued e.g. that "common preferences are always possible". A common preference is an outcome which all sides prefer to their initial preference – wholeheartedly with no regrets, downsides, compromises or sacrifices. It's strictly better than alternatives, not better on balance.

In BoI, DD writes about problems being soluble – and what he means by solutions is strictly win/win solutions which satisfy all sides in this sort of arbitration.

An arbitration tool is new ideas (which are usually small modifications of previous ideas). For example, take one side's idea but modify a few parts to no longer conflict with what the other side wants.

As long as every side wants good things, there is a solution like this to be found. Good things don't inherently conflict.

Sometimes sides want bad things. This can either be an honest mistake, or they can be evil or irrational.

If it's an honest mistake, the solution is criticism. Point out why it seems good but is actually bad. Point out how they misunderstood the implications and it won't work as intended. Or point out a contradiction between it and something good they value. Or point out an internal contradiction. Analyze it in pieces and explain why some parts are bad, but how the legitimate good parts can be saved. When people make honest mistakes, and the mistake is pointed out, they can change their mind (usually only partially, in cases where only part of what they were saying was mistaken).

How can a side be satisfied by a criticism/refutation? Why would a side want to change its mind? Because of explanations. A good criticism points out a mistake of some kind and explains what's bad about it. So the side can be like, "Oh, I understand why that's bad now, I don't want that anymore." Good arguments offer something better and make it accessible to the other side, so they can see it's (strictly) better and change their mind with zero regrets (conflict actually resolved).

If there is an evil or irrational mistake, things can go wrong. Short answer: you can't arbitrate for sides which don't want solutions. You can't resolve conflicts

with people who want conflict. Rational epistemology doesn't work for people/sides/ideas who don't want to think rationally. But one must be very careful to avoid declaring one's opponents irrational and becoming an authoritarian. This is a big issue, but I won't discuss it here.

Arbitration ends when there's exactly one win/win idea which all sides prefer over any other options. There are then no (relevant to the issue) conflicts of ideas. (DD would say no "active" conflicts). Put another way, there's one non-refuted idea.

Arbitration is a creative process. It involves things like brainstorming new ideas and criticizing mistakes. Creative processes are unpredictable. A solution could take a while. While a solution is *possible*, what if you don't think of it?

Reasonable sides in the arbitration can understand resource limits and lower expectations when arbitration resources (like time and creative energy) run low. They can prefer this, because it's the objectively best thing to do. No reasonable party to an arbitration wants it to take forever or past some deadline (like if you're deciding what to do on Friday, you have to decide by Friday).

When the sides in a conflict are different people, the basic answer is the more arbitration gets stuck, the less they should try to interact. If you can't figure out how to interact for mutual benefit, go your separate ways and leave each other alone.

With a conflict between ideas in one person, it's trickier because they can't disengage. One basic fact is it's a mistake to prefer anything that would prevent a solution (within available resources) – kind of like wanting the impossible. The full details of always succeeding in these arbitrations, within resource limits, are a big topic that I won't include here.

How do justificationists handle arbitrations? They hear each side and add and subtract points. They tally up the final scores and then declare a winner. The primary reason the loser gets for losing is "because you scored fewer points in the discussion". The loser is unsatisfied, still disagrees, and there's still a conflict, so the arbitration failed.

Here's a different way to look at it. Each side in arbitration tries to explain why its proposal is ideal. If it can persuade the other side, the conflict is resolved, we're done. If it can't, the rational approach is to treat this failure to persuade as

"huh, I guess I need better ideas/explanations" not as "I have the truth, but the other guy just won't listen!"

In other words, if either side has enough knowledge to resolve the conflict, then the conflict can be resolved with that knowledge. If neither side has that, then both sides should recognize their ideas aren't good enough. Both sides are refuted and a new idea is needed. (And while brilliant new ideas to solve things are hard to come by, ideas meeting lowered expectations related to resource limits are easier to create. And it gets easier in proportion to how limited resources are, basically because it's a mistake to want the impossible.)

Justificationism sees this differently. It will try to pick a winner from the existing sides, even when (as I see it) they aren't good enough. As I see it, if the existing sides don't already offer a solution (and only a fully win/win outcome is a solution), then the only possible way to get a solution is to create a new idea. And if any side doesn't like it (setting aside evil, irrationality, not wanting a solution, etc), then it isn't a solution, and no amount of justifying how great it is could change that.

To relate this back to some of the original topics:

The arbitration model doesn't involve confidence levels or probabilities. Ideas have boolean status as either win/win solutions (non-refuted), or not (refuted), rather than a score or rank on a continuum. Solutions are explanations – they explain what the solution is, how it solves the problem(s), what mistakes are in all attempted criticisms of this solution, why it's a mistake to want anything (relevant) that this solution doesn't offer, why the things the solution does offer should be wanted, and so on. Explanation is what makes everything work and be appealing and allows conflicts to be resolved.

Final Comments

I don't expect you to understand or agree with all of this. Perhaps not much, I don't know. To discuss hard issues well requires a lot of back-and-forth to clear up misunderstandings, answer questions and objections, etc. Understanding has to be created iteratively (Popper would say "gradually" or "piecemeal").

I am open to discussing these topics. I am open to considering that I may be wrong. I wouldn't want a discussion to assume a conclusion from the start. I tried

to explain enough to give some initial indication of what my epistemology is like, and some perspective about where I'm coming from.

Footnotes

[1]

My point was, whatever your method for preserving bodies, you could assign it some odds, arbitrarily. You could say cremation causes less damage than shooting bodies into the sun, so it has better revival odds. And then pick a small number for a probability. You need to have an argument regarding vitrification that couldn't be said by someone arguing for cremation, burial or freezing.

There should be something to clearly, qualitatively differentiate cryonics from alternatives like cremation. Like it should differentiate vitrification not as better than cremation to some vague degree, but as actually on a different side of an reasonably explained might-work/doesn't-work line.

Here's an example of how I might argue for cryonics using scientific research.

Come up with a measure of brain damage (hard) which can be measured for both living and dead people. Come up with a measure of functionality or intelligence for living people with brain damage (hard). Find living brain damaged people and measure them. Try to work out a bound, e.g. people with X or less brain damage (according to this measure of damage) can still think OK, remember who they are, etc.

Vitrify some brains or substitutes and measure damage after a suitable time period. Compare the damage to X.

Measure damage numbers for freezing, burial and cremation too, for comparison. Show how those methods cause more than X damage, but vitrification causes less than X damage. Or maybe the empirical results come out a different way.

Be aware that when doing all this, I was using many explanations as unconscious assumptions, background knowledge, explicit premises, and so on. Expose every part of this stuff to criticism, and for each criticism write an explanation addressing it or modify my view.

Then someone would be in a position to make a non-arbitrary claim favorable to cryonics.

This is not the only acceptable method, it's one example. If you could come up with some other method to get some useful answers, that's fine. You can try whatever method you want, and the only judge is criticism.

But something I object to is assigning probabilities, or any kind of evaluations, without a clear method and explanation of it. (E.g. where does your 10% for cryo come from? Where does anyone's positive evaluation come from?)

I don't think it's reasonable for Alcor or CI to ask people to pay 5-6 figures without first having a good idea about how to judge today's cryonics (like my example method). And from a decision making perspective, I expect people asking for lots of money – and saying they can perform a long term service for me in a reliable way – should have some basic competence and reasonable explanations about their stuff. But instead they put this on their website:

http://www.alcor.org/Library/html/CaseForWholeBody.html

It offers a variation on Pascal's Wager to argue for full-body cryo over neuro (basically, get full body just in case it's necessary for cryo to work). No comment is made on whether we should also believe in God due to Pascal's Wager. And it states:

> Now, what if we would relax our assumptions a little and allow for some degree of ischemia or brain damage during cryopreservation? It strikes us that this further strengthens the case for whole body cryopreservation because the rest of the body could be used to infer information about the non-damaged state of the brain, an option not available to neuropatients.

No. I'm guessing you also disagree with this quote, so I won't argue unless you ask.

There are some complications like maybe Alcor is confused but today's cryonics works anyway. I won't go into that now.

[2]

We can, whenever we want, create ranking systems which we think will be useful for some purpose (somewhat like defining new units of measurement, or defining new categories to categorize stuff with).

The judge of these inventions is criticism. E.g. someone might criticize a ranking system by pointing out why it isn't effective for its intended purpose.

Concretely, we could rank body preservation methods by the amount of brain damage after 10 years. Then, in that system, we'd rank vitrification > freezing > burial > cremation.

Whether this is useful depends on context (which Popper calls the problem situation). What problem(s) are we trying to solve? Do we have a non-refuted idea for how to use the ranking in any solutions?

Our example ranking system has some relevance to people who consider brain damage important, but not to people who believe the goal should be to preserve the soul by using the most holy methods. They'd want to rank by holiness, and might rank vitrification last.

This is important because the rankings only matter in the context of some explanations of how they matter and for what (which must deal with criticism).

So ranking is secondary to explanation. It can't come first. This makes ranking unsuited for dealing with epistemology issues such as how to decide which explanations to accept in the first place.

In summary, we can make something up, argue why it's effective for a purpose, and if our argument is successful then we can use it for that purpose. This works with rankings and many other things.

But this is different than epistemology rankings, like trying to rank how good ideas are, or how probable, or how high quality of explanations they are.

Or put another way: to rank *those* things, you would have to specify how that ranking system worked, and explain why the results are useful for what. That's been tried a lot. I don't think those attempts have succeeded, or can succeed.

Aubrey de Grey Discussion, 6

> In a nutshell, I think most of what you’rve written here comes down to something I already entirely agree with, namely that any kind of ranking of competing ideas is inferior to the identification of a win-win. You don’t need to persuade me of that.

ok but i have stronger claims:

1) All human choices can and should be made using the win-win arbitration approach. it is the only method of rational thinking

2) Justificationism doesn't work at all, and has zero value as an alternative method

> My preferred way of looking at this is that identifying a win/win is the extreme case of choosing by ranking, in rather the same sense that Popperian decision-making is the limiting case of Bayesian decision-making. But I mention that only for clarification; if you think it’s wrong, do tell me, but let’s not spend too much time on that (not yet anyway) because I don’t think it affects the rest of what I want to say.

I don't agree that Bayesian epistemology has any value. OK I won't argue that now. Though FYI DD's latest blog post is "Simple refutation of the ‘Bayesian’ philosophy of science":

http://www.daviddeutsch.org.uk/2014/08/simple-refutation-of-the-bayesian-philosophy-of-science/

> My problem comes down to the impracticality of the arbitration approach. I can certainly believe that all conflicts can reliably be resolved in bounded time, but as you say, we have the problem of needing to make a decision now (or soon), not in 10000 years.

I'm glad to hear that. It's a big point of agreement. Most people think some problems aren't solvable, and some human conflicts don't have any possible win/win outcomes.

I meant the arbitration approach can always be done within real life time limits. Or at least scenarios where you have some time to think. For a starting point, let's limit discussion to cases where the time limit is at least an hour. And definitely not worry about the 5 millisecond case.

In Oct 2002, I made a similar objection to yours. **DD answered** why common preference finding (a.k.a. win/win arbitration) doesn't require infinite creativity.

> ... the finding of a common preference does not entail finding the solution to any particular problem.
>
> The economy does not require infinite creativity to grow. Particular enterprises fail all the time. Particular inefficiencies may remain unimproved for long periods. The economy as a whole may have brief hitches where mistakes have been made and have to be undone; but if it stagnates to the extent of failing to innovate, there is a reason. It's not just 'one of those things'. The reason has nothing to do with there being a glut of nautiluses on the market, but is invariably caused by someone (usually governments, but in primitive societies also parents) forcibly preventing people from responding to market forces. Stagnation is not a natural state in a capitalist economy; it has to be caused by force.
>
> Science does not require infinite creativity to make new discoveries. Particular lines of research fail all the time but where science as a whole has ceased to innovate it is never because the whole scientific community has turned its attention to the nautilus but invariably because someone (governments and/or parents) has forcibly prevented people from behaving according to the canons of scientific rationality.
>
> An individual personality does not require infinite creativity to grow. Particular a priori wants go unmet all the time, and large projects also fail and sometimes a person has a major life setback. But if they get stuck to the extent of failing to innovate it is not because they have spontaneously wandered into a state where their head resembles a nautilus but because someone has forcibly thwarted them once (or usually a thousand times) too often.

> ...
>
> ... problems can be continually solved without infinite creativity, without perfect rationality, and without relying on any particular problem being solved by any particular time. And that is sufficient for -- in fact it is what *constitutes* -- economic growth, scientific progress, and human happiness.

Why DD equates innovation with win/win arbitration isn't explained here. One way to understand it is because we consider win/win arbitration to be the only epistemological method capable of creating knowledge, solving problems, making progress/innovation, etc.

The point that problem solving (or conflict resolution) in general doesn't require solving any particular problem is very important. That's what allows fast solutions.

What you can do is ask questions in arbitration like, "Given we think we won't solve problems X, Y and Z within our resource constraints, what should we do?" That question can be answered without solving problems X, Y or Z, and its answer can be a successful win/win arbitration outcome.

As with everything, it's open to criticism, e.g. a side might think X actually can be solved within the resource constraints. Then all sides might be able to agree, for example, to try to solve X, but also to set up a backup plan in case that doesn't work.

If an arbitration seems particularly hard relative to the resources available, a longer exclusion list can be proposed. By setting things aside as necessary, arbitration can succeed in the short term.

I also have a bunch of writing on this topic. E.g.:

http://fallibleideas.com/avoiding-coercion

And I gathered multiple links at:

http://curi.us/1595-rationally-resolving-conflicts-of-ideas

There's a lot. One reasonable way to approach this is read things until you find a specific point of disagreement or two, then comment. Maybe just the material in this email is enough.

I'm providing the links partly so if you like reading it, it's available to you. But if you prefer a more back-and-forth approach, that's fine with me. I like writing.

> So we should absolutely put some effort into looking for a resolution, but the amount of effort we should put in before we throw in the towel and retreat to ranking is a trade-off between our commitment to making the right choice and our urgency to make any choice. Just as in life generally, in fact! - and that’s no accident, because even though it seems much more informal, all the decisions we make in life are subject to the same epistemic logic concerning science that you are setting out. The perfect is the enemy of the good, and all that.

I didn't intend to limit stuff to science. Yes, epistemology applies to the whole of life.

I would say more like, "wanting the impossible is an enemy of the good". But I'd be cautious because people often underestimate what's possible (e.g. with SENS).

> Moreover, it turna out that the arbitration approach is considerably more impractical in some areas than in others, and biology is a particularly impractical one - basically because the complexity of the system under discussion and the depth of our ignorance of its details lead to the arising of lots of very similarly-ranked (by Occam’s razor, for example) conflicting ideas.
>
> In a way, it seems to me that you’re describing arbitration rather in the way that mathematics works. A mathematical proof is (so I’m told, and it makes sense to me) no more nor less than an argument that other mathematicians find persuasive.

I agree about math being fallible and thinking of math proofs as arguments.

Lots of people think math proofs are infallible. DD criticized that in The Fabric of Reality.

> So the discussion of a proposed proof is a process of arbitration between the belief that the conjecture is open and the belief that it is resolved (say, that it is true). And we find that mathematics lies at the opposite extreme from biology in terms of practicality: mathematicians tend to be able to agree really quite quickly whether a candidate proof holds water.

To be clear about my stronger claims above: I don't think which field affects arbitration practicality, since it always works.

> So, let’s look at your cryonics proposal:
>
> > Here's an example of how I might argue for cryonics using scientific research.
> >
> > Come up with a measure of brain damage (hard) which can be measured for both living and dead people. Come up with a measure of functionality or intelligence for living people with brain damage (hard). Find living brain damaged people and measure them. Try to work out a bound, e.g. people with X or less brain damage (according to this measure of damage) can still think OK, remember who they are, etc.
> >
> > Vitrify some brains or substitutes and measure damage after a suitable time period. Compare the damage to X.
> >
> > Measure damage numbers for freezing, burial and cremation too, for comparison. Show how those methods cause more than X damage, but vitrification causes less than X damage. Or maybe the empirical results come out a different way.
>
> I would assert that you makes my case extremely well. Consider your first two steps, coming up with these measures. It’s actually really easy to come up with such measures - lots and lots of alternative ones.

It's easy to come up with bad measures. For good measures, I'm not convinced.

Part of my perspective on this has to do with how bad IQ tests and school tests are, and the great difficulty of doing better.

> The only way to decide which to use is to (gasp) rank them, according to your third step, testing their correlation with function.

That isn't the only way. You could come up with an explanation of what measure you should use, and why, and expose it to criticism.

It's very important to consider explanations. E.g. percentage of undamaged brain cells could be tried in a measure because we have an explanatory understanding that more undamaged cells is better. And we might modify the measure due to the locations of damaged cells, because we have some explanatory understanding about what different region[s] of the brain do and which regions are most important. It'd be a mistake to try arbitrary things as a measure and then look for correlations.

Typical correlation approaches are bad science because they are explanationless. If one does have explanation, that explanation should be primary. An explanation can reference a correlation and explain why it matters, and only then would a correlation matter.

Correlation is [a] big topic. I think we should focus more on arbitration. But here's an initial explanation of correlation related problems.

Summary: explanationless correlation approaches to science are the same kind of thing as induction.

There are infinitely many correlations out there. What people do is find and focus on a small number of correlations, and pay selective attention to those.

The only thing that can make this selective attention reasonable is an explanation. And it should be a clear, explicit explanation that's exposed to criticism, not an unstated one that secretly governs which correlations get attention.

I think about this in a more general way which might be helpful. A correlation is a type of pattern. There are infinitely many patterns in the world you could find, most meaningless, and they only matter when there's an explanation that they do.

And there's also the problem that if you find a sequence, e.g. "2,2,2,2,2" and you think it's a pattern, you actually have no knowledge of how it will continue unless you have an explanation. Which brings us to induction, because dealing with sequences like this and say "oh it's going to be 2 next" – without an explanation – is a major inductivist activity. If the sequence is over time, the inductivist might add, "the future is likely to resemble the past".

Similarly if you find X correlates with Y during a particular time period, the assumption they will continue to correlate in a different future time period – without explanation – is basically "the future is likely to resemble the past", a.k.a. induction.

Selective attention is also a feature of induction. Inductivists look at evidence and notice it's consistent with several ideas of interest to them. But don't pay serious attention to the infinitely many other ideas that evidence is equally consistent with. And some of those ignored ideas, which are equally "supported" by the evidence, contradict the ideas getting their selective attention.

A further issue is that context matters. You can only understand what would be a significant change in circumstances (such that one wouldn't expect a correlation or pattern to continue) via an explanatory understanding of what context is relevant and what would be a significant change.

On a related note, suppose a ranking system is developed for something, and even assume it's good. How do you know if it's still applicable when dealing with anything that isn't absolutely literally 100% identical to the original context? How do you know which changes matter? How do you know if which country you're in is part of the relevant context that can't be changed? How do you know if the calendar year is part of the relevant context that can't be changed? Only by explanation. Only by understanding why the ranking system works can you tell what changes would mess that up and what changes wouldn't.

And how can you judge explanations and decide which ones are good? The win/win arbitration method.

> Er, but there are loads of ways to test function too, so any such ranking (even setting asite the precision of measurement and such like) is only finitely reliable. The rest of what you say would be fine if we really could come up with a way to define and then measure brain damage that was

> unequivocally 100% reliable - but unfortunately, in the real world with the time we have, we can't do that. So, we have no choice but to survey our various options for the measure of damage and function and the measurability of those measures, rank them according to something or other, and make our decision as to whether cryonics is worth doing on that basis - but, do so using some probability threshold of how likely we need it to be to work in order to justify the expense, so as to incorporate our uncertainty as to whether we have measured the brain damage correctly and accurately. If we can't successfully perform your first steps, we have no right to proceed as if we had performed all steps - which is precisely what you're doing by rejecting the (admittedly inferior, but doable) ranking approach and just subjectively saying you don't think the available data justify spending that much money.
>
> Tell me what's wrong with the above.

Regarding rankings, they are OK when you have an explanation of why a particular ranking system will get you a good answer for a particular problem. In other words, deciding to use that ranking system for that purpose is the outcome of a win/win arbitration. If you don't have that, rankings are arbitrary.

The rankings could be fully arbitrary. Or they could have some reasons, but arbitrarily ignore some criticism or problem. (If no criticism or problem was being irrationally ignored, then it would be a win/win arbitration outcome). Another common approach to rankings is to intentionally design the ranking system so it reaches a predetermined conclusion which people already think is plausible not arbitrary.

My main point here is that if they haven't done my proposal, they should have done something else with an explanation of why it makes sense. They have do something, have some explanation, some knowledge.

They actually do have basic explanations, e.g. I've read one of them saying that vitrified brains look pretty OK, not badly damaged, to the unaided human eye. The implication is damage that's hard to see is small, so cryopreservation works well. This is a bad argument, but it's the right type of thing. They need this type of thing, but better, before anyone should sign up.

I think you have in your mind some explanations of the right type, but haven't said them because of your methodology that doesn't emphasize explanation as I do. So I don't know how good they are.

In footnote [1], I comment on a couple cyro papers and information about fracturing.

I also have a second way for judging Alcor and CI specifically. Consider the explanation, "Preserving people for much later revival is a very hard problem. Hard problems like this don't get solved by accident by irrational and incompetent methods, they require things like scientific or intellectual rigor."

As usual, one can't explain everything at once. This explanation leads to further questions like why people don't accidentally solve hard problems. An important thing about explanation, persuasion and win/win arbitration is you only have to satisfy objections that any side cares to make, not all possible objections. If no one thinks an objection is good, don't worry about it. Yes you could miss something important, but there are always infinitely many possible objections and you can't answer all of them, you have to go by the best knowledge anyone has of which are important, and if mistakes are made due to ignorance, so be it, that's not always avoidable.

(Explanations sometimes answer infinite categories of objections. But to answer literally all possible objections would basically require omniscience

Another aspect I didn't explain here is how incompetent and irrational Alcor and CI are. But I did give an initial explanation of that previously. And I have in my mind more extensive explanation of it, if you raised objections to my initial explanation.

A reason bad people don't solve hard problems is because mistakes and problems are inevitable, so there has to be rational problem solving and mistake-correcting taking place or else advanced stuff will never work. Since I don't see Alcor and CI doing a decent job with that, I don't think their service works.

---

[1]

http://198.170.115.106/reports/Scientific_Justification.pdf

> A rabbit kidney has been vitrified, cooled to -135C, re-warmed and transplanted into a rabbit.

Rabbit was fine. Cool.

> When cooling from -130C to -196C thermal stress on large solid vitrified samples can cause cracking and fracturing.

But rabbit kidney was not cooled to the relevant colder temperatures. This has footnote 27.

> Due to its more well-defined nature, cracking damage may be much easier to repair than freezing damage.

This is too vague, plus doesn't say anything about how much damage there is. It has no footnote. Paper lacks better information than this about fracturing damage issues.

Footnote 27:

http://www.sciencedirect.com/science/article/pii/0011224090900386

One of their main conclusions given in the abstract:

> fracturing depends strongly on cooling rate and thermal uniformity

So one question one might have is: what cooling rates do Alcor and CI use? How much thermal uniformity do they achieve? But to my knowledge they don't carefully measure that kind of information, or even use sufficiently standardized procedures to get consistent results.

Also kind of scary, the 2008 paper is citing information from 1989, rather than more recent information.

Another paper:
http://www.lorentzcenter.nl/lc/web/2012/512/problems/4/Long-

term%20storage%20of%20tissues%20by%20cryopreservation.pdf

This one has lots of interesting information about why cryonics is hard, and ends by saying, "In summary, we hope to have demonstrated that tissue cryopreservation is a complex problem..." The article can give one a sense of how hard these problems are, and therefore why it takes scientific rigor, top quality knowledge and rational problem-solving ability to succeed at human cryonics. Which Alcor and CI lack.

There's also some information about how bad vitrification damage is here:

http://lesswrong.com/lw/343/suspended_animation_inc_accused_of_incompe

It's from an expert and I've found no contrary information. Example statements:

> There is no present technology for preserving people in a "fairly pristine state" at cryogenic temperatures. Present cryopreservation technology even under perfect conditions causes biological effects such as toxicity and fracturing that are far more damaging than the types of problems you've expressed concern about.
>
> ...
>
> Most cryobiologists would regard the idea of repairing organs that had cracked along fracture planes as preposterous, as I'm sure you do if you believe that 300 mmHg arterial pressure or one hour of ischemia is fatal to a cryonics patient.

In that first quote, we get an actual comparison of vitrification damage to something else. That something else is, "the types of problems you've expressed concern about". Those problems are, from the parent comment:

> a bunch of unqualified, overgrown adolescents, who want to play doctor with dead people, while pretending to be surgeons and perfusionists

In summary, Brian Wowk (an expert on Alcor's board of directors) is saying that damage from vitrification, without any errors by cryo personnel, is "far more damaging" than the various horror stories of gross error by cryo personnel. And far more damaging than, e.g., an hour of ischemia.

I'm no expert on this, but trying to look it up, it seems a few minutes of ischemia causes brain damage. And there are explanations for this, e.g. "central neurons have a near-exclusive dependence on glucose as an energy substrate, and brain stores of glucose or glycogen are limited" [2]. Damage far worse than an hour of ischemia sounds to me like cryo's not going to work yet, and I haven't found information to the contrary.

[2] **http://www.ncbi.nlm.nih.gov/pmc/articles/PMC381398/**

Aubrey de Grey Discussion, 7

> You’re telling me that that’s not the right way to make a decision, but I’m still not seeing the details of the alternative approach you recommend. Can you please spell it out in similar terms - specifically, in terms that make clear how it can be performed in a chosen amount of time (say, a week)?

This can't be answered completely directly because part of the point is to think about epistemology in a different way. Creative thinking does not follow a specific formula. (Or at least, the formula is complicated enough we don't know all the exact details – or we'd have AGI already.)

Making decisions requires creative thought. The structure of creative thought is: solve problems using the method of guesses and criticism, which leads to a new situation with new problems.

(Guesses and criticism is the only method which creates knowledge. It's literally evolution, which is the only solution ever figured out to the problem of creating knowledge. I'm hoping you have some familiarity with this already from Popper and DD, or I could go into more detail.)

This structure is not a series of steps to be done in order. For example, guesses come before criticism to have something to criticize, but also afterwards to figure out how to deal with the criticism. And criticisms are themselves guesses. And criticisms need their own criticism to find and improve mistakes, or they'll be dumb.

And as one works on this, his understanding of the problem may improve. At which point he's in a new situation which may raise new problems already, before the original problem is resolved.

One can list options like, in response to criticism of a guess: revise understanding of that guess, make brand new alternative guesses, adjust the existing guess not to be refuted, criticize the criticism, or revise understanding of the problem.

But there's no flowchart saying which to do, when. One does one's best. One thinks and uses judgment. But some methods are bad and there's criticisms of them.

The important thing, like Popper explained about democracy, is not so much what one is doing right now, but if and how effectively mistakes are being found and improved.

Everyone has to start where they are. Use the best judgment one has. But improve it, and keep improving it. It's progress that's key. Methods shouldn't be static. Keep a lookout for problems, anything unsatisfactory, and then make adjustments. If that's hard, it's OK, exchange criticism with others whose set of blind spots and mistakes doesn't exactly overlap with one's own.

What if one misses something? That's why it's important to be open to discussion and to have some ways for ideas from the public to reach you. So if anyone doesn't miss it, you can find out. (http://fallibleideas.com/paths-forward) What if everyone misses something? It can happen. Actually it does happen, routinely. There's nothing to be done but accept one's fallibility and keep trying to improve. Continual progress, forever, is the only good lifestyle.

While there isn't a rigid structure or flowchart to epistemology, there is some structure. And there are some good tips. And there are a bunch of criticisms that one should be familiar with and then not doing anything they refute.

The win/win arbitration model provides a starting point with some structure. People have an idea of how arbitration works. And they have an idea of how a win/win outcome differs from a compromise or win/lose outcome.

Internal to the arbitration, creative thought (which means guesses and criticism) must be used. How do arbitrations end in time? Participants identify the problem that it might not, guess how to finish in time, and improve those ideas with criticism. That is, in a pretty fundamental way, the basic answer to everything. Whatever the problem is, guess at the solution and improve the guesses with criticism.

This raises questions like:

- what if one can't think of any guesses for something?

- what if one has some bad guesses, but can't think of any criticisms?
- what if one has several guesses and gets stuck deciding between them?
- what if different sides in an arbitration disagree strongly and get stuck?
- what if no one has any ideas for what would be a win/win solution?
- what if the sides in the arbitration keep fighting instead of discussing rationally
- what if the arbitration runs into resource limits?
- what if there is one or more issues no one has an answer to, how can arbitration work around those?

Rather than a flowchart, epistemology offers answers to all of these questions. Does that make sense? Would you agree that the loose method above, plus answers to all questions like this (and all criticisms) would be sufficient and satisfactory?

If you agree with the approach of addressing those questions (plus you can add some), and it would persuade you, then I'll do that next. Part of the reason the discussion is tricky is because we're starting with different ideas of what the goalposts should be.

I would also like to give more in the way of concrete examples but that's very hard. I can tell you why it's hard and try some examples.

People use these methods, successfully, hundreds of times per day. They get win/win solutions in mental arbitrations, routinely. Most of these are individual, and some are in small groups, and it isn't routine in large groups.

Examples of these come off as trivial. I'll give some soon.

People also get stuck sometimes. And what they really want are examples of how to solve the problems they find hard, get stuck on, and are irrational about. But I can't provide one-size-fits-all generic examples that address whatever individual readers are stuck on. And even if only talking to one person, I'd have

to find out what their problems are, and solve them, to provide the desired examples.

If I wasn't concerned about privacy, I could give examples of problems that I had a hard time with, and solved. But it wouldn't do any good. People will predictably react by thinking my solution wouldn't work for them because they are different (true), or that problem I struggled with was always easy for them (common), or knowing my solution to my problem won't solve their problems (true).

Here are some examples of routine win/win arbitrations:

Guy is hungry but doesn't want to miss TV show. Decides to hit pause. Solved. (Other people would grab some food during a commercial. The important thing is the person doing it fully prefers it for their life.)

People want to eat together, but want different types of food. Go to a food court with multiple restaurants. Solved.

Person wants to buy something but hesitates to part with their money. Thinks about how awesome it would be, changes mind, happily buys. Solved.

Person wants to buy something but hesitates to part with their money. Estimates the value and decides it's not actually worth it. Changing mind about wanting it, happily doesn't buy. Solved.

Person wants to find their keys so they can leave the house, but doesn't feel like searching. Thinks about how great the sushi will be, finds he now wants to search for the keys, does so happily. Solved.

Person wants to get somewhere in car but is in unwanted traffic, some part of his personality wants to get mad. He thinks about how getting mad won't help, doesn't get mad.

All life is creative problem solving, and people do it routinely. And people change their mind about things, even emotions, routinely, in a win/win way without regrets or compromise. But people don't find these examples convincing, because they see these examples as unlike whatever they find hard and therefore notable. Or they find some of these hard, e.g. they hate looking for their keys, or have "road rage" problems.

Here's a more complex hypothetical example.

I want to borrow my child's book, which is in the living room, but he's not home. I have conflicting ideas about wanting the book now, but not wanting to disturb his things. While I want to respect his property, that doesn't feel concretely important, so I'm not immediately satisfied. I resolve this by remembering he specifically asked me never to disturb his things after a previous mistake. I don't want to violate that, so I change my attitude and am concretely satisfied that I shouldn't borrow his book, and I'm happy with this result.

I go on to brainstorm what to do instead. I could read a different book. I could buy the ebook from Amazon instantly (many people would consider this absurd, but books are very very cheap compared to the value of getting along slightly more smoothly with one's family). I could write an email instead of reading. I could phone my kid and ask permission.

Here is where examples can get tricky. Which of those solutions do I do? Whichever one I'm happy with. It depends on the exact details of my ideas and preferences. But whichever option works for me might not work so well for a reader imagining themselves in a similar situation. Their problem situation is different than mine, and needs its own creative problem solving applied to it.

And what if I don't like any of these options, can't think of more, and get stuck? Well, WHY? There is some reason I'm getting stuck, and there is information about what the problem is and why I'm stuck. What I should do depends on why I'm stuck. And why you would be stuck in a similar situation won't be the same as why I got stuck. You won't identify with my way of getting stuck, nor with what solutions work to get me unstuck.

So, I decide that phoning is easy, and I don't like giving up without trying when trying is cheap. So I phone.

9/10 times in similar situations with similarly reasonable requests, kid says yes. This time, kid says no.

9/10 scenarios kinda like this where kid says no, I HAPPILY accept this and move on to figuring out what else to do. This is easy to be happy to go along with because I respect (classical) liberal values, and I know there are great options available in life which don't violate them, so I'm not losing out.

1/10 times, I tell my kid how I'm really eager to read the book, and there's no electronic version for sale.

Then, 9/10 times, kid says "oh ok, then go ahead". 1/10s he still says no.

If he still says no, 9/10 I accept it because I care about respecting his preferences for his property, and I have plenty of alternative ways to have a good day. I want both a good day and to respect his property, and I can have both. And I don't want to be pushy and intrude on his life over something minor – it's not even worth the transaction costs of making a big deal out of – so I won't.

And 1/10 times I say "i'm sorry to bug you about this, but i ran out of stuff to do and was actually kinda sad, and then i thought of this one thing i wanted to do, which is read this book, and i got excited, and i'm really dreading going back to my problem of being bored and sad. so, please? what's the big downside to you?"

And then 9/10 times kid agrees, but 1/10 times he says "still no, sorry, but i wrote private notes in the margins of that book, do not open it".

And the pattern continues, but additional steps get exponentially rarer. The pattern is that at each step, usually one finds a way to prefer that outcome, and sometimes one doesn't and continues. Note at each step how it's harder to continue asking, it takes more unusual reasons.

DD persuaded me of the rule of thumb that approximately 90% of interpersonal conflicts, dealt with rationally, get resolved per step trying to resolve. I know this isn't intuitive in a world where people routinely fight with their families.

If you disagree, it's not so important. If someone's methods are wrong, and it causes any problems, and someone else knows better, that's no big deal. Methods can be criticized and changed. Correct or not, the approach in the example is – like many others – just fine as a starting point.

All of life can and should go smoothly with problem solving and progress. It often doesn't because of irrationality, because of not understanding the right epistemology, because of bad values, because of anti-rational memes, because of deeply destructive parenting and education practices. All of those are solvable problems which change people's intuitions about what lifestyles work, but which do not change what epistemology is true.

As a final example, let's take cryonics. Here is something I can say about it: I have given some arguments which you have not criticized and I have not found refutations for anywhere else. On the other hand, if you tell me any arguments against my position, I will either refute ALL of them or change my mind in some way to reach an uncriticized position. (Note refuting includes not just saying why the argument is false, but also for example why it's true but doesn't actually contradict my position.)

You create a 10% estimate in a vague way, which you describe as a subjective estimate of a feeling. This hides your actual reasoning, whatever it is, from criticism – not just criticism by me but also by yourself.

You gather arguments on all sides, but you don't analyze them individually and judge what's true or not and why. I do. That is a very key thing – to actually go through the arguments and sort out what's right and wrong, to learn things, to figure the subject out. It's only by doing that, not just kinda making up an intuitive conclusion, that progress and problem solving happen.

You see the situation as many arguments on both sides and want a method for how to turn those many arguments into one conclusion.

I see the situation as many arguments, which can be analyzed and dealt with. Many are false, and one can look through them and figure things out. My current position is that literally every known pro-cryonics-signup argument is false in the context of my situation, and most people's situations.

(Context is always a big deal. People in different situations can correctly reach different conclusions specific to their situation. For example a rich person with a strongly pro-cryonics wife might find signing up increases marital harmony, and has no downsides that bother him, even though he doesn't believe it can work.)

It's this critical analysis of the specific arguments by which one learns, by which progress happens, etc. It always comes down to critical challenges: no matter how great some side seems, if there is a criticism of it, that criticism is a challenge that must be answered, not in any way glossed over.

If the criticism cannot be refuted (today), one must change his mind to something no longer incompatible with the point (pending potential new ideas). It's completely irrational and destructive of problem solving to carry on with any idea which has any criticism one can't address.

There are many ways to deal with criticisms one can't directly refute. And these methods are themselves open to criticism. We could talk more about how to do this. But the key point is, any method which doesn't do this is very bad. Such as justificationism, and the specific version of it you outlined, which allow for acting contrary to outstanding unanswered criticisms.

> The first may be only a point of clarification. While I certainly agree that we rationally choose which correlations to pay attention to on the basis of explanations, I think we have a problem that those explanations themselves emerge from analysis of other correlations, which were paid attention to because of other explanations, and so on, right back to correlations that we arbitrarily decide we don't need to explain, such as that every time we measure the fundamental physical constants we get the same answers. This seems to me to tell us that explanations can't be viewed as inherently better than correlations - they are part and parcel of a single process, just as science proceeds by an alternation between hypothesis formation and hypothesis testing. What am I missing?

Explanations come from brainstormed guesses in relation to problems. (And are improved with criticism for error-correction, or else the quality will be awful.)

There is no process which starts with correlations and outputs explanations (or more generally, knowledge).

Most correlations are due to coincidence. They are not important.

A correlation matters when referred to in an explanation. It has no special interest otherwise. Just like dust particles, blades of grass, mosquitos, copper atoms. There's dust all over the place, most is not important, but some can be when mentioned in an explanation.

The issue of getting started with learning is not serious, because it doesn't really matter where one starts. Start somewhere and then make improvements. The important thing is the process of improvement, not the starting point. One can start with bad guesses, which are not hard to come by.

Also we do have an explanation of why different experiments measuring the speed of light in a vacuum get the same answer. Because they measure the same thing. Just like different experiments measuring the size of my hand get the same

answer. No big deal. The very concepts of different photons all being light, and of them all having the same speed, are explanatory ideas which make better sense out of the underlying reality.

> The second one is possibly also just something I'm misunderstanding. For any pioneering technology that we have not yet perfected - SENS, cryonics, whatever - there are always explanations for why it is feasible (or, in the case of cryonics, why part of has already been achieved even though we won't know that for sure until the rest of it also has) and other explanations for why it isn't. I think what you're saying is that the correct thing to do is to debate these explanations and eventually come up with an agreed winner, and that in the meantime the correct thing to do is to triage, by debating explanations for what we should do in the absence of an agreed winner between the first set of explanations, and act on the basis of an agreed winner between that second set of explanations. But I don't see how that can work in practice, because the second debate will typically come down to the same issues as the first debate, so it will take just as long. No?

A second debate on the topic, "given the context of issues X, Y, Z being unresolved, now what?" cannot come down to the same issues as the first debate, because they're specifically excluded.

It may be helpful to look at it in terms of what IS known. Part of the context is people do know some things about SENS, cryo, or whatever topic. So there is an issue of, given that known stuff, what does it make sense to do about it?

When discussions get stuck in practice, it's not because of ignorance. If no one knows X yet, that doesn't make two people disagree, since that's the same for both of them, it's a point in common. The causes of disagreements between people are things like irrationality or different background knowledge like values or goals; perhaps someone has a lifetime of tangled thinking that's hard to sort out. The solution to those things are (classical) liberal values like tolerance, individualism, leaving people alone, and only interacting for mutual (self-perceived) benefit.

Take for example:

http://www2.technologyreview.com/sens/

The reason those debates didn't resolve your differences is because those people directed their creativity towards attacking SENS, not truth-seeking. Rational epistemology only works for people who choose to use it. The debate format was also deeply unsuited to making progress because it allowed very little back-and-forth to ask questions and clear up misunderstandings. It wasn't set up for creating mutual understanding, none of your opponents wanted to understand SENS, the results were predictable, but that has nothing to do with what's possible. (BTW, awful as this sounds, it isn't such a big deal, since they aren't going to use violence against you. Not even close. So you can just go on with SENS and work together with some better people.)

BTW notice the key thing about that debate: you could answer all of their criticisms. ALL. Specifically, not vaguely.

And I think you know that if you couldn't, that'd be a serious problem for SENS.

Take the claim, "even though these [SENS] categories are sometimes so general as to be almost meaningless, they still omit many age-related changes that contribute to senescence, including age-related increases in oxidative damage and changes in gene expression."

If you had no answer to that, SENS would be in trouble. It only takes one criticism to refute something. But you had the answer. And not in some vague way like, "I feel SENS is 10% likely to work, down from 20% before hearing that argument". But specifically you had an actual answer that makes the entire difference between SENS being refuted and SENS coming out completely fine.

This is a good example of how things can actually get resolved in debates. Like the claim about oxidative damage, that can be resolved, you knew how to resolve it. Progress can be made, things can be figured out. (Though not for those who aren't doing truth-seeking.)

Challenges like the oxidative damage argument can routinely be answered and discussions can resolve things. What you said should have worked. It only didn't because the other guy was not using anything resembling a rational epistemology, and did not want progress in the discussion.

> The third one is where I’m really hanging up, though. You say a lot about good and bad explanations, but for the life of me I can’t find anything in

> what you've said that explains how you're deciding (or are claiming people should decide) HOW good an explanation needs to be to justify a particular course of action.

Answer: that is the wrong question.

There is no such thing as how epistemologically good an explanation is.

The way to judge explanations I'm proposing is: refuted or non-refuted. Is there a criticism pointing out any flaw whatsoever? Yes or no?

No criticism doesn't justify anything. It just makes more sense to act on ideas with no known flaws (non-refuted) over ideas with known flaws (refuted).

One common concern is criticisms pointing out minor flaws, e.g. a typo, or that a wording is unclear. The answer is: if the criticism really is minor, then it will be easy to fix, so fix it. Create a new idea (a slight modification of the old idea) to which the criticism doesn't apply.

Or explain why a particular thing that seems like a flaw in some vague general way is not a flaw in this specific context (problem situation). Meaning: it seems "bad" in some way, but it won't prevent this approach from working and solving the problem in question.

For example, someone might say, "It'd be nice if the instruments on the space shuttle were 1000x more accurate. It's bad to have inaccurate instruments. That's my criticism." But a space shuttle has limited finite goals, it's not supposed to be perfect and do everything, it's only supposed to do specific things such as bring supplies to the space station, land on the moon, or complete specific experiments. Whatever the particular mission is, if it can be completed with the less accurate instruments, then the "inaccurate instruments are bad" criticism doesn't apply.

> In the case of cryonics, you've read a bit about where the practice of cryonics is today and you've come to the conclusion that it doesn't currently justify signing up, because you prefer the arguments that say the preservation isn't good enough to the ones that say it is. But you don't say where the analysis process should stop.

Stop when there is exactly one non-refuted idea. I am unaware of any non-refuted criticisms of my position on the matter.

This has nothing to do with preferring some arguments. I am literally unaware (despite looking) of any argument to sign up with Alcor or CI, that I can't refute right now today. (Though as I mentioned above, I have in mind my situation or most situations, but not all people's situations. In unusual situations, unusual actions can make sense.)

In your method you talk about gathering arguments for both sides. I have tried to do that for cryonics, but I've been unable to find any arguments on the pro-cryonics side that survive criticism. Why do you think give it a 10% chance to work? What are any arguments? And meanwhile I've given arguments against signing up which you have not individually, specifically refuted. E.g. the one about organizations that are bad at things don't solve hard problems because problems are inevitable so without ongoing problem solving it won't work.

I think a lot of the reason debates get stuck is specifically because of justificationist epistemology. People don't feel the need to give specific arguments and criticisms. Instead they do things like create arbitrary justification/solidity/goodness scores that are incapable of resolving the disagreements between the ideas.

> For example, you say:
>
> > percentage of undamaged brain cells could be tried in a measure because we have an explanatory understanding that more undamaged cells is better. And we might modify the measure due to the locations of damaged cells, because we have some explanatory understanding about what different region of the brain do and which regions are most important.
>
> We might, yes, or we might not. How do you decide whether to do so?

Creative thinking. Guess whether it's a good idea and why. Improve this understanding with criticism.

> And if you decide that we should take account of location, why stop there? Suppose that someone has proposed a reason why neurons with more synaptic connections to other neurons matter more. It might be a really really hand-wavey explanation, something totally abstract concerning the holographic nature of memory for instance, but it might be consistent with available data and it might also be really hard to falsify by experiment.

Almost all refutation is by argument, not experiment. (See: section about grass cure for the cold in FoR, where DD explains that even most ideas which are empirical and could be dealt with by experiment, still aren't).

Since you call it "hand-wavey", what you mean is you have a criticism of it. The thing to do is state the criticism more clearly, and challenge the idea: either it answers the criticism or it gets thrown out.

> So, should we take it into account and modify our measure of damage accordingly? What's worse, we don't even know whether we have even heard all the relevant explanations that have been proposed, even ignoring all the ones that will be proposed in the future. There might be ones that we don't know that conflict with the ones we do know, and that we might eventually decide are better than the ones we do know. Shouldn't we be taking account of that possibility somehow?

Yes. One should make reasonable efforts to find out about more ideas, and not to block off other people telling one ideas (http://fallibleideas.com/paths-forward).

You will ask what's reasonable, how much is enough. Answer: creative thinking on that point. Guess what's the right amount of effort to put into these things (given limits like resource constraints) and refine the guess with some critical thinking until it seems unproblematic to one. Then, be open to criticism about this guess from others, and try to notice if things aren't going well and one should reconsider.

> This seems to bring one inexorably back to the probabilistic approach. Spelling it out in more detail, the probabilistic approach seems to me to consist of the following steps:

> - Gather, as best one can in the time one has decided to spend, all the arguments recommending either of the alternative courses of action (such as, sign up with Alcor or don’t);
>
> - Subjectively estimate how solid the two sets of arguments feel;

How? This vague step hides a thousand problems in its details.

> - Estimate how often scientific consensus has, in the past, changed its mind between explanations that initially were felt to differ in solidity by that kind of amount, and how often it hasn’t (with some kind of weighting for how long the prevailing has been around);

This has a "future will resemble the past" element without a clear explanation of what will be the same and what context it depends on.

And it glosses over the details of what happened in the various cases, and the explanations of why.

It also gives far too much attention to majority opinion rather than substantive arguments.

It's also deeply hostile to large innovations in early stages. Those frequently start with a large majority disagreeing and feeling the case for the innovation has very low solidity.

If you look at the raw odds that a new idea is a brilliant innovation, they suck. There are more ways to be wrong than right. You need more specific categories like, "new ideas which no one has any non-refuted criticism of" – those turn out valuable at much higher rates.

> - Use that as one’s estimate of one’s likelihood of being right that the seemingly more solid of the two sets of explanations is indeed the correct set, hence that the course of action that that set recommends is the correct course;
>
> - decide what probability cutoffs motivate each of the three possible ways forward (sign up and focus on something else until some new item of data is brought to one’s attention, don’t sign up and focus on something else

> until some new item of data is brought to one’s attention, or decide to spend more time now on the question than one previously wanted to), and act accordingly.

This approach involves no open-ended creative thinking and not actually answering many specific criticisms and arguments. Nor does it come up with an explanation of the best way to proceed. It does not create knowledge.

This proposed justificationist method does not even try to resolve conflicts between ideas. It doesn't try to figure out what's right, what's wrong, or why. There's no part where anything gets figured out, anything gets solved, anyone learns anything about reality. It's kind of like a backup plan, "What if rational thinking fails? What if progress halts? Under that constraint, what could we do?" Which is a bad question. It's never a good idea to use irrational methods as a plan B when rational methods struggle.

One of the weirder things about discussing justificationism is, I know you frequently don't use the method you propose. It's only to the extent that you don't use this method that you get anywhere. Like at **http://www2.technologyreview.com/sens/**

You didn't present your subjective feeling of the solidity of SENS, or estimates about how often a scientific consensus has been right, or anything like that. You did not gather all the anti-SENS arguments and then estimate their solidity and give them undeserved partial credit without figuring out which are true and which false. Instead, you gave specific and meaningful arguments, including refuting ALL their criticisms of SENS. Then you concluded in favor of SENS not on balance – you didn't approach it that way – but because the pro-SENS view is the one and only non-refuted option available for answering the debate topic.

Aubrey de Grey Discussion, 8

Thanks again Elliot. I have several issues below, but they have a single common theme.

> This approach involves no open-ended creative thinking and not actually answering many specific criticisms and arguments. Nor does it come up with an explanation of the best way to proceed. It does not create knowledge.

I was probably unclear on that: that's part (most, in fact, for interesting cases) of step 1, i.e. "Gather, as best one can in the time one has decided to spend, all the arguments recommending either of the alternative courses of action." I didn't mean to imply that this would be restricted to pre-existing arguments. So in other words, yes actually, I did use exactly this method in my evaluation of Estep's criticism of SENS, and in my reply I articulated some of the results of that evaluation, namely some refutations of elements of the criticism. Consider your position as a reader: why did you accept my rebuttal as the last word? Why didn't you write to Estep to ask him for a more thorough re-rebuttal than TR gave him the option of? Answer (I claim): because you subjectively decided that my rebuttal was impressive ENOUGH that Estep PROBABLY wouldn't have a persuasive re-rebuttal, so you chose not to allocate time to contacting him. Note the quantitative, as well as subjective, elements of what I claim was your process (and I claim it confidently, because I can't think of any other process you could have used for deciding not to write to Estep).

It's interesting you specifically express confidence, and can't think of any other process. This description isn't close to how I approached the Estep debate.

First, your rebuttal wasn't important here. I had already decided Estep was wrong before reading your rebuttal. That was easy. His position was largely philosophy, rather than being about detailed scientific points that I might have difficulty evaluating. While reading his text, I thought of criticisms of his arguments.

Actually, rather than being particularly impressed, I disliked three aspects of your rebuttal. But these criticisms were tangents, and are standard parts of academic culture. If I'm right about them, they don't make SENS wrong or Estep right. 1) Complaining about Estep's invective and saying you'd take the high road, but then returning some invective. 2) What I consider an overly prestigious writing style, partly intended to impress. 3) Arguing some over who has how much scientific authority and what they think (rather than only discussing substantive issues directly).

My interest in your rebuttal wasn't to learn why Estep was wrong – which I already knew. Note I say why he was wrong (explanation) rather than considering who is more impressive (ugh). Instead, I read to see how closely your thinking and approach matched my own (if I found important differences, I'd be interested in why, at least one of us would have to be wrong in an important way), to see what passes for debate in these kinds of papers in your field, and to see if you'd say an important point I'd missed or a mistake.

The main reason I didn't write to Estep is because I don't think he wants to have a discussion with me. My usual policy is not to write to paper authors who don't include contact information in their papers.

Now that you brought it up, I tried google and didn't find contact info there either. I think discussion is unwelcome. I did find his email in the GRG archives, but that's no invitation.

I actually would be happy to talk to him, if he wanted to have a discussion. Like if Estep volunteered to answer questions and criticisms from me, I'd participate. I like to talk to a variety of people, even ones I consider very bad. I want to understand irrationality and psychology better. And it helps keep my ideas exposed to all kinds of criticism. And I don't get myself stuck in unwanted polite or boring conversation.

You're right that I wouldn't expect Estep to change my mind if we talked. This is because I guessed an understanding of what he's like, which I have no criticisms of and no non-refuted alternatives to. Not probability. But this is minor. I'd talk to him anyway, the issue is he doesn't want to.

And I didn't just leave this to my judgment. I exposed my view on this matter to criticism. I wrote about it in public and invited criticism from the best thinkers

I've been able to gather (or anyone else). (BTW you'd be welcome to join my Fallible Ideas discussion group and my private group.)

I don't do more than this because I have explanations of why other activities are better to spend my time on, and I don't know a problem/criticism with my approach or an explanation of a better approach. And all of this is open to public criticism. And I've made a large ongoing effort to have ready access to high quality criticism.

> > There is no such thing as how epistemologically good an explanation is.
>
> I don’t get this. You’ve been referring to good and bad explanations throughout this exchange. What have you been meaning by that, if not epistemologically good and bad? I know you are saying that there are only refuted or non-refuted explanations, but you must have been meaning something else by good and bad, since you’ve definitely been using those adjectives - and other ones, like “clear”, “explicit” etc - in an unambiguously quantitative rather than binary/boolean sense, e.g.:

I can see how that'd be confusing. It's an imprecise but convenient way to speak. Depending what you're doing, you only need limited precision, so it can be OK. And it'd take forever to elaborate on every point, it's better only to go into detail on points where someone thinks it's worthwhile to, for some reason.

My position is that all correct arguments can be converted or translated into more precise statements that strictly adhere to the boolean epistemology approach.

Speaking of amount of clarity is a high level concept that's sometimes precise enough. You can, when you want to, get into more precise lower level details like pointing out specific ambiguous phrases or unanswered questions about the writer's position.

Saying an explanation is good or bad (in some amount) can quickly communicate an approximate evaluation without covering the details. It's loose speaking rather than epistemology.

> > They actually do have basic explanations, e.g. I've read one of them saying that vitrified brains look pretty OK, not badly damaged, to the unaided human eye. The implication is damage that's hard to see is small, so cryopreservation works well. This is a bad argument, but it's the right type of thing. They need this type of thing, but better, before anyone should sign up.
>
> If it’s the right type of thing, what’s “bad" about it?

It is the right type of thing, meaning: it involves explanation and argument.

"Bad" here was an imprecise way to refer to some arguments I didn't write out upfront.

Damage that's hard to see to the naked human eye is not "small" in the relevant sense. The argument is a trick where it gets people to accept the damage is small (physical size in irrelevant regular daily life context), and implies the damage is small (brain still works well).

Why use unaided human eye instead of microscope? It's a parochial approach going after the emotional appeal of what people can see at scale they are used to. Rather than note appearances can be deceiving and try to help the reader understand the underlying reality, it tries to exploit the deceptiveness of appearances.

And it doesn't attempt to explore issues like how much damage would have what consequences. But with no concept of what damage has what consequences, even a correct statement of the damage wouldn't get you anywhere in terms of understanding the consequences. (And it's the consequences like having one's mind still revivable, or being dead, that people care about.)

> - and more to the point, how bad?

Refuted.

> What is your argument for saying "They need this type of thing, but BETTER (quantitative…), before anyone should sign up”? How much better, and why?

It needs to be better to the point it isn't refuted. Because it's a bad idea to act on ideas with known flaws.

(There are some complications here like they don't actually know my criticism, the flaws aren't known to them. What is "refuted" in each person's judgment depends on their individual knowledge. That's a tangent I won't write about now.)

> You can’t just say “non-refuted”, because you know as well as I do that any argument about anything interesting can be met with a counter-argument, which itself can be met, etc., unless one has decided in advance how to terminate the exchange.

No, I disagree!

It's hard to keep up meaningful criticism for long.

Yes someone can repeat "That's dumb, I disagree" forever. But a criticism, as I mean it, is an explanation of a flaw/mistake with something, and this kind of bad repetitive objection doesn't explain any mistakes.

I don't think you had this kind of repetition in mind, or you wouldn't have specified "about anything interesting". "That's dumb, I disagree" can be used on trivial topics just as well as interesting topics.

I think you're saying that substantive critical discussion doesn't terminate and keeps having good points indefinitely. Until you terminate it arbitrarily.

I think good points are hard to come by. What are "good" points here, specifically? Ones which aren't already refuted by pre-existing criticism.

As you go along in productive discussions, you build up criticisms of many things. Not just of specific points, but of whole categories of points. Some of the criticisms have "reach" as DD calls it. They have some level of generality, they

apply to many things. As criticism builds up, it gets progressively harder to come up with new ideas which aren't already refuted by existing criticism.

The reason many discussions don't look like this in practice is because of irrationality and bad methods, rather than discussions having to be that way.

> My fundamental problem remains: you haven’t given me a decision-making algorithm that terminates, or even usually terminates, in an amount of time that I can specify in advance.

It's a mistake to 100% rigidly specify time limits in advance. Reasoning for time limits should be open to criticism.

The closest to a flowchart I can give you is something like:

- think creatively etc, as discussed previously
- when nearing a resource limit (like time), start referring to this limit in arguments, to bring arbitration to a close. e.g. instead of "I disagree with that, and here's why in detail", a side might say, "I disagree with that, but we don't have time to get into it. Instead, here is what I propose that we may both find acceptable."
- as resources get tighter, it gets easier to please all sides. like, they may agree it's better to flip a coin than not to reach a decision by a certain deadline.
- reasonable sides understand their fallibility and don't want anyone to go along with something without persuasion. and they understand persuasion on some point can exceed a resource limit. so they actively PREFER to find mutually agreeable temporary measures for now, when appropriate, while working on persuasion more in the longer term as more resources are available
- sometimes things go smoothly. no problem. sometimes they don't. when they don't, there are specific techniques which can be used.
- specifically, one considers questions of the form, "Given the context - and specifically not reaching agreement on points X, Y and Z, but having

agreement on A, B and C - what can be done that's mutually agreeable? What can be done on this issue with the limited agreement?"

- while working on this new question, if there are any sticking points, then a similar question can be asked adding those sticking points to the exclusion list.

- these questions reduce the complexity and difficulty of the arbitration as low as needed.

- the more you use questions like this and temporarily exclude things due to resource limits, the easier it is to reach agreement. if it's different people, it goes to "since we disagree so much, let's go our separate ways". the harder case is either when a person has conflicting ideas or two people are entangled (e.g. parent and child). but that still reaches outcomes like, "given we disagree so much, and we need a decision now, let's flip a coin". both sides can prefer that to any known alternatives, in which case it's a win/win outcome.

- but what if they don't agree to flip a coin over it? well, why not? this is fundamentally why a flowchart doesn't work. because people disagree about things for reasons, and you can't flowchart answers to those reasons.

- but basically sides will either agree to a coin flip (or some better default they know of), or else they will propose something they consider a better idea. a better idea while being reasonable – so like, something they think the other side could agree with, not something that'd take a great deal of persuasion involving currently-unavailable resources.

- if sides are unreasonable – e.g. try to sabotage things, or just want their initial preference no matter what – then any conflict resolution procedure can stall or fail. that's unavoidable.

- this doesn't terminate in predictable-in-advance time because sometimes everyone agrees that the deadline is less important than further arbitration, and prefers to allocate more resources. i don't think this is a problem. it can terminate quickly when that's a good idea. the only reason it won't terminate quickly is specifically because a side disagrees that terminating quickly is a good idea in this case. (and if that happens, there will be a reason in

> context, which may be right or wrong, and there is no one-size-fits-all flowchart answer to it, it matters what the reason is)

I have one. It's not perfect - I accept all your criticisms of it, I think - but the single feature that it terminates in a reasonably predictable time (just how predictable is determined, of course, by how close together one chooses the two cutoff probabilities to be) is so important that I think the method is better than any alternative that doesn't reliably terminate.

The thing is, I think you DO have an algorithm that reliably terminates, and that despite your protestations it is pretty much identical to mine. Look at this example for illustration:

> Also we do have an explanation of why different experiments measuring the speed of light in a vacuum get the same answer. Because they measure the same thing. Just like different experiments measuring the size of my hand get the same answer. No big deal. The very concepts of different photons all being light, and of them all having the same speed, are explanatory ideas which make better sense out of the underlying reality.

Nonsense, because each measurement measures different photons, and we have no better explanation for all photons having the same speed than for all pigeons having the same mass. This is not trivial: indeed, I recall that Wheeler made quite a big deal out of the awfully similar question of the mass of the electron and proposed that there is in fact only one electron in the Universe. We have explicitly made the choice not to enquire further on the question.

If you go deeper, then yes I don't know everything about physics. There's some initial explanations about this stuff, but it's limited.

I'm unclear on why this is important. I don't study physics more because I prefer to do other things and I don't know of any criticisms/problems with my approach. Even if I did study physics all day, I still wouldn't know everything about it and would make choices about which things to enquire further about, because I couldn't do everything at once. I would think of an explanation for

how I should approach the matter, adjust or rethink until no criticism, and do that.

> Or this one:
>
> > Person wants to buy something but hesitates to part with their money. Thinks about how awesome it would be, changes mind, happily buys. Solved.
>
> That only works with an additional step that comes just before “happily buys”, namely “switches brain off before remembering that one might soon change one’s mind back”. And, actually, another step that says “remembers that one is really good at not crying over spilt milk, i.e. once the money is spent one is happy to live with whatever regret one might later have”. And so on. I know you know this.

But I don't know it. I deny it.

I think switching off the brain and trying not to think of some issues, because one couldn't deal with the issues if he paid attention to them, is a really bad approach. It's choosing winners in an irrational way – instead of resolving the conflict of ideas, you're playing the role of an arbiter who only lets one side speak, then declares them the winner.

About spilt milk: Sometimes people think of that and it helps them happily buy something. But sometimes people don't. It's not required. There are many optional steps that people find useful, or not, depending on their specific circumstances.

> But, yet, you were fine with just writing “Solved”! I conclude that you DO have a termination procedure in your algorithm, and moreover that it’s an indisputably vague and subjective and probabilistic and epistemologically hole-riddled one just like mine, and I don’t know why you’re having such trouble admitting it.

I don't concede because I disagree.

I think a rational non-hole-riddled epistemology is possible, and that I understand it.

> Let’s get back to cryonics - largely because I am now somewhat invested in the goal of changing your mind about signing up, coupled of course with the equally legitimate converse goal of giving you a fair shot at changing mine.
>
> Let’s start with the specific question I already referred to above:
>
> > They actually do have basic explanations, e.g. I've read one of them saying that vitrified brains look pretty OK, not badly damaged, to the unaided human eye. The implication is damage that's hard to see is small, so cryopreservation works well. This is a bad argument, but it's the right type of thing. They need this type of thing, but better, before anyone should sign up.
>
> As this stands, as I just said, it is too vague to be amenable to refutation even in principle, i.e. it doesn’t meet your own epistemological standards, because it doesn’t incorporate any statement of (let alone any argument for) your criterion for how good that explanation needs to become.

my standard is: is there a criticism of it? not some criterion for how good.

> As above, “non-refuted” doesn’t work, because that relies on consideration of (for example) how much time I choose to allocate to giving you refutations and how much you choose to allocate to giving me refutations, and I sense that that that’s a decidedly non-level playing field.

You mean, it's not a level playing field because I allocate more time to trying to get this issue right? Or at least to writing down my thinking, so that if I'm mistaken someone could tell me?

BTW, what is your explanation of why no one has written good explanations of why to sign up for cryonics anywhere? Why have they left it to you to write it, instead of merely link things?

(Good explanations to what standard? Your own. If stuff met your standards you'd link it instead of writing your own.)

> My (unashamedly justificationist) starting-point is that the absence of gross damage feels like enough evidence for revivability to satisfy me that people should sign up.

The evidence you refer to is consistent with infinitely many positions, including ones that conclude not to sign up for cryo. Considering it evidence for a specific conclusion, instead of others it's equally consistent with, is some mix of 1) arbitrary 2) using unstated reasons

Why should a fact fully compatible with non-revivability be counted as "evidence for revivability"?

> So let’s start with you amplifying your above statement, with a sense of what you WOULD view as a good enough (yes I said it) argument, to give me some goalposts to aim for.

The goalposts fundamentally are: I don't have further criticism.

This is hard because I have many criticisms. But there really have to be ways for me to get answers to all of them (though not all from you personally). Or else you'd be asking me to do something I have a reason not to do; you'd be asking me to just ignore my own judgment arbitrarily for no reason.

I also think you overestimate how problematic this is because you're used to debates that don't go anywhere, don't resolve anything, because of how terribly irrational most people are.

Another big factor is people who don't want to be persuaded. Rational persuasion is impossible with unwilling subjects. People always have to persuade themselves and fill in lots of details, you can't tell them everything and perfectly customize it all to their context and integrate it with all their other ideas. They have to play an active role, or any persuasion will be superficial.

Something that I'd see as a good starting place is explanations connecting different amounts of damage to consequences like being fine or dead, and

quantifying the amount of damage Alcor and CI cause today.

Aubrey de Grey Discussion, 9

> Thanks. Hm. I’m sincerely trying my very hardest to understand what you’re saying about your own thought processes, but I’m not making much progress.

I understand. It's very hard. Neither DD nor Popper had much success explaining these things in their books. I mean the books are great, but hardly anyone has thoroughly been persuaded by those books that e.g. justificationism is false.

I'm trying to explain better than they did, but that's tough. It's something I've been working on for a long time, but I haven't yet figured out a way to do it dramatically more effectively than DD and Popper. I think correct epistemology is very important, so I keep working at it. But I'm not blaming you or losing patience or anything like that.

> At this point I think where I’m getting stuck is that the differences between your and my descriptions of how you make decisions (and of how one ought to make decisions) mainly hinge on the distinction between (a) not having any further criticisms and (b) not choosing to spend further time coming up with further criticisms,

I think there's a misunderstanding here.

I wouldn't draw a distinction there. If you don't know more criticisms, and resolved all the conflicts of ideas you know about, you're done, you resolved things. Whether you could potentially create more criticisms doesn't change that.

The important thing is not to ignore (or act against) any criticisms (or ideas) that you do know about. Either ones you came up with, or someone told you.

If you do know about a conflict between two ideas, don't arbitrarily pick a side. Rationality requires you either resolve the conflict, or proceed in a way that's neutral regarding the unresolved conflict. This is always possible.

Does that summarize one of my big points more clearly?

In other words, when there's a disagreement, either figure out how to resolve it or how to work around it, but don't assume a conclusion while the debate is ongoing. (The relevant ongoing debate typically being the one in your own mind. This isn't a formula to let irrational confused people hold you up indefinitely. But details of how to deal with this aspect are complex and tricky.)

Secondarily it's also important to be open to criticism and new ideas. If the reason you don't know about a criticism is you buried your head in the sand, that's not OK. (This part is pretty uncontroversial as an ideal, though people often don't live up to it very well.)

> and I claim that for most interesting questions that is a distinction that is very hard to make, because it’s almost always fairly easy to come up with a new criticism (and I don’t mean a content-free one like “that’s dumb”, I mean a substantive one). Now, you disagree - you say "It's hard to keep up meaningful criticism for long”. That’s absolutely not my experience. In fact I would go further: I think that the way our brains work is that exhaustion or distraction from what we objectively know we’d like to do is a phenomenon that we generally like to put out of our minds, because we wish it weren’t so, so it’s virtually impossible to know whether we have truly exhausted our potential supply of criticisms. I really, really like to know why I think what I think, so I feel I go further down these rabbit-holes than most people, but they’re still rabbit-holes.

I'm mainly concerned with actual criticisms and conflicts of ideas, not potential.

Apart from the issue of willfully not thinking of arguments you couldn't answer, or choosing not to hear them, then it's only the actual ideas you have that matter and need conflict resolution between them now.

> I think the only promising-sounding way to resolve this (i.e. to determine how difficult it really is to keep up meaningful criticism - which will very probably entail gaining a better understanding of each other’s threshold of “meaningful”) is for us to work through a concrete example. Naturally I suggest we continue with cryonics.

I disagree with "only". But that's fine, sure.

Though, actually, I don't think cryonics is ideally suited because on cryonics I'm more in the role of critic, and you more in the role of defending against criticism.

But our epistemology disagreement is kind of along the lines of: I have higher standards. So when I'm in the role of critic, this will come off as: my criticism is picky and demands standards you think can't be met.

If we used a different topic where I have a lot of knowledge and positive claims exposed to criticism, it could more easily be you making criticisms as picky as you want – trying to demonstrate such picky criticisms can't be answered – and then me showing how to answer them.

What do you think?

I reply about cryonics below anyway.

> Before that, though, I have a new issue with some of what you said in this latest reply. You seem to have created a massive loophole in your approach here:
>
> > - the more you use questions like this and temporarily exclude things due to resource limits, the easier it is to reach agreement. if it's different people, it goes to "since we disagree so much, let's go our separate ways".
>
> I can’t for the life of me see how you can seriously view that as an epistemologically acceptable outcome. And yet, I claim that it is indeed necessary to say that in order to reach your claim that resource limitations are not fatal to the epistemologically respectable method you advocate. Agreeing to disagree is no different from saying “that’s dumb”, except insofar as the participants may have gained a better understanding of the issues (negligibly better, in most cases, I claim). This is particularly important because of the non-level-playing field issue - much more often than not, the two participants in a debate will have unequal resource limits, so one of them will need to quit before the other feels ready to quit, so going separate ways ends up as the only option.

I'm unclear on the problem. If people AGREE to leave each other alone, and act accordingly, then they have a mutually agreeable win/win outcome that neither of them has a criticism of. This resolves the conflict between them that they were trying to sort out.

This doesn't resolve the tough problems in the field – but they know that and aren't claiming otherwise. What their agreement resolves is the problems surrounding their immediate decision making about how to deal with each other.

> OK, let’s get back to cryonics.
>
> > BTW, what is your explanation of why no one has written good explanations of why to sign up for cryonics anywhere? Why have they left it to you to write it, instead of merely link things?
>
> I think what’s been written by Alcor is (in aggregate) a good explanation, and you’ve read it already, so I didn’t suggest you read it.

In aggregate, I think you will agree it contains flaws. I've pointed some out.

So what's needed to save it is some modifications. Some way to have a position similar to it, without the flaws.

But I've been unable to figure out a position like that. And I haven't found Alcor's material to be much help for doing this.

I'm also unclear on what you think the gist of Alcor's case is. What primary claims make up their argument that you think is good? I actually have very little concept of what you think their website says.

Do you think their website presents something like your argument below? That's not what I got from it.

> > The evidence you refer to is consistent with infinitely many positions, including ones that conclude not to sign up for cryo. Considering it

> > evidence for a specific conclusion, instead of others it's equally consistent with, is some mix of 1) arbitrary 2) using unstated reasons
>
> Why should a fact fully compatible with non-revivability be counted as "evidence for revivability"?

> In most scientific fields, and certainly in almost all of biology, the totality of available evidence is consistent with infinitely many positions, including the position that eating grass cures the common cold.

yes

> Thus, one doesn’t reject the position that eating grass cures the common cold on the basis of a boolean approach to available evidence - one does so on the basis, as you said, that the quality of explanations for why eating grass cures the common cold (i.e. refutations of the position that eating grasss does not cure the common cold) is inadequate - there are no “meaningful” such explanations.

i disagree and think one should approach the grass-cures-cold with specific criticisms, not vague quality/justification judgments. Examples below.

> Let’s have a go. Grass contains huge numbers of phytochemicals that we have identified, and the limitations of breadth and depth of our investigations are such that we can be quite sure it also contains lots that we have not identified. Phytochemicals have many diverse properties, such as antioxidant properties, that are shared with compounds that are known to have therapeutic effects on the common cold. Kids occasionally eat grass, and they occasionally recover faster than average from the common cold, so in order to know whether grass cures the common cold we would need to survey the cases of this to determine whether the two were positively correlated, and no one has done this. I don’t claim that this is a meaningful refutation of the position that eating grass doesn’t cure the common cold, but I do claim that it is a meaningful refutation of the position that it’s not worth doing the experiment to determine whether eating grass cures the common cold. I don’t claim that it’s a persuasive refutation, but the only reason I have for distinguishing between persuasive and meaningful is

> probabilistic/justificationist: based on my subjective intuition, I think the chances of the experiment coming out on the side that grass indeed cures the common cold are too low to justify the resources needed to do the experiment. What am I missing?

This argument is fine in the sense of being unlike "that's dumb" with no reason given. It's "meaningful". To put it approximately but perhaps communicate effectively: I wasn't trying to exclude anything even 1% as reasonable as this.

But this passage makes several mistakes. Here are some criticisms:

It's suggesting resources be allocated to this. But it doesn't compare the value it thinks can be gained by this change in resource allocation to the value gained from current allocation. So it doesn't really actually argue its case and is vague about what specifically should be done.

It's too much of a "try this, it might work" approach. There are more promising leads. One way (of many) to get more promising leads is to think of a specific mechanism by which something could work which you don't know how to rule out given current evidence and arguments, and then test that.

Another mistake is looking for correlation itself, when the thing we actually care about is causation (we care whether eating grass CAUSES recovery from colds). A good project would try to determine causation. This could maybe involve looking at correlations, but there'd have to be an idea about what to usefully do with the correlation information if found.

Note BTW that all three of these criticisms use fairly general purpose ideas. They're mildly adapted from previous discussions of other topics. For that reason, it doesn't take much work to create them. And as one builds up a greater knowledge of general purpose criticisms, it gets harder to propose any ideas that pass initial criticism using already-known criticism techniques.

> Back to cryonics.
>
> > Damage that's hard to see to the naked human eye is not "small" in the relevant sense. The argument is a trick where it gets people to

> > accept the damage is small (physical size in irrelevant regular daily life context), and implies the damage is small (brain still works well).
> >
> > Why use unaided human eye instead of microscope? It's a parochial approach going after the emotional appeal of what people can see at scale they are used to. Rather than note appearances can be deceiving and try to help the reader understand the underlying reality, it tries to exploit the deceptiveness of appearances.
> >
> > And it doesn't attempt to explore issues like how much damage would have what consequences. But with no concept of what damage has what consequences, even a correct statement of the damage wouldn't get you anywhere in terms of understanding the consequences. (And it's the consequences like having one's mind still revivable, or being dead, that people care about.)
>
> Sure, all agreed - but they are not making that mistake. It's known that living systems have pretty impressive self-repair machinery, and that it tends to work better to repair physically smaller damage than physically larger damage. Therefore, even though we know perfectly well that damage too physically small to be seen with the naked eye could still be too much for revivability, we know that there is a whole category of damage that would indeed (probably) be too much and is absent,

ok

> and that's meaningful evidence.

Meaningful evidence – meaning what?

This evidence is consistent with many things, so if you want to bring it up you should give an explanation about what it means. It doesn't speak for itself.

Do you mean that of the infinitely many cryo-doesn't-work possibilities, an infinite subset have been ruled out? Yes. Do you mean that this raises the amount of remaining cryo-does-work possibilities relative to the cryo-doesn't-work possibilities? No, infinity doesn't work that way.

> Plus, of course Alcor (and more importantly 21CM) have looked at vitrified tissue with microscopes and not seen appreciable damage

What do you mean "appreciable" and where do they provide this information? Aren't fractures appreciable damage?

How does this fit with Brian Wowk's comments, brought up earlier, about lots of damage? Do you think he was mistaken, or is this somehow compatible?

> - but how much magnification is enough? If they were basing everything on 100X microscopic images, what would be your procedure for deciding whether or not to complain that they hadn’t looked at the EM level?

I'd ask WHY they didn't use EM level and see if I see something wrong with their answer. There ought to be an explanation, presumably already written down.

I'd hope the answer wasn't "lack of funds even though it's very important". That'd be a plausible but disappointing answer I could imagine getting.

Not using the best microscopes around would strike me as suspicious enough to ask a question about. But in that scenario, I wouldn't be surprised to find they had a reason I have no criticism of, and then I'd drop it. Advanced technology sometimes has drawbacks in some cases, rather than being universally the best option.

> I can certainly provide (as Alcor do) positive evidence for how much damage is tolerable - but of course there are ways to refute it, but only if one views one’s refutations as meaningful. For example, we can look at the amount of variabiity in structure of the brain in non-demented elderly, and we can see big differences between people who are equally cognitively healthy - easily big enough to be seen without a microscope.

Damage and non-damage variation are different things. What is this comparison supposed to accomplish?

People have different ideas. It would unsurprising if this has significant physical consequences since ideas have to have physical form. Though we also can see non-microscopic differences in healthy hearts, lungs, skin, etc, so the easily visible brain differences don't necessarily mean more than those other differences.

> You could say, ah, but all one is doing there is identifying changes that are not harmful - but that’s circular, in the absence of direct evidence as to whether the damage done by vitrification is harmful.

I'm unclear what you're saying would be circular, or how you'd answer my comments in the section right above. I think I didn't quite get your point here, unless my comments above address it.

To phrase this as a direct criticism, for the context of me being persuaded, the issues have to be clear to me, so things I find unclear won't work.

To succeed in this context, they have to be either modified to be clear to me (which I always try to do myself before objecting), or else there'd have to be auxiliary explanations, either about the specific subject, or about how to read and think better, so that I could then get the point.

> Is that a refutation that you would view as meaningful? If so, what’s your re-refutation of it? And if not, why not?

Yes, meaningful. I think the bar there is real low. I just wanted to exclude complete non-engagement like a tape recorder could accomplish.

Some answers above. Plus this doesn't address some points I raised previously, but we can set those aside for now.

Aubrey de Grey Discussion, 10

> I wouldn't draw a distinction there. If you don't know more criticisms, and resolved all the conflicts of ideas you know about, you're done, you resolved things. Whether you could potentially create more criticisms doesn't change that.

OK, of everything you've said so far that is the one that I find least able to accept. Thinking of things takes time - you aren't disputing that. So, if at a given instant I have resolved all the conflicts I know about, but some of what I now think is really really new and I know I haven't tried to refute it, how on earth can I be "done"?

As you say, you already know that you should make some effort to think critically about new ideas. So, you already have an idea that conflicts with the idea to declare yourself done immediately.

If you know a reason not to do something, that's an idea that conflicts with it.

> That's precisely what I previously called switching one's brain off. Until one has given one's brain a reasonable amount of time to come up with a refutation of a new concept, the debate is abundantly ongoing.
>
> You make a good point about the cryonics example being sub-optimal because I'm the defender and you're the critic. So, OK, let's do as you suggest and switch (for now) to a topic where you're the defender and I'm the critic. There is a readily available one: your approach to the formation of conclusions.

I see some problems with this choice:

Using an epistemology discussion as an example for itself adds complexity.

Using a topic where we disagree mixes demonstrating answering criticism with trying to persuade you.

Using a complex and large topic is harder.

I still will criticize justificationism because you still think it can create knowledge.

If I were to pick, I'd look for a simpler topic where we agree. For example, we both believe that death from aging and illness is bad. If SENS or cryonics succeeded, that would be a good thing not a bad thing.

I wonder if you think there's criticisms of this position which you don't have a refutation of? Some things you had to gloss over as "weak" arguments, rather than answer?

The idea that grass cures the common cold – or that this is a promising lead which should be studied in the near term – would also work. You gave an initial argument on this topic, but I replied criticizing it. You didn't then demonstrate your claimed ability to keep up arguments for a bad position indefinitely.

> (Does it have a name?

Popper named it Critical Rationalism (CR).

> - presumably something better than non-justificationism? I’m going to call it Elliotism for now, and my contrary position Aubreyism, since I have a feeling we’re both adopting positions that are special cases of whatever isms might already have been coined.) Let’s evaluate the validity of Elliotism using Elliotism.

What do you mean by "validity"? I'm guessing you mean justification.

To evaluate CR with CR, you would have to look at it with its own concepts like non-refutedness.

> The present state of affairs is that I view Elliotism as incorrect - I think justificationism is flawed in an ideal world with infinite resources (especially time) but is all we have in the real world, whereas (as I understand it) Elliotism says that justificationism can be avoided and a purely boolean approach to refutation adopted, even in a resource-constrained world.

Yes, but, I think you've rejected or not understood important criticism of justificationism. You've tried to concede some points while not accepting their conclusions. So to clarify:

Justificationism is not a flawed but somewhat useful approach. It literally doesn't and can't create knowledge. All progress in all fields has come from other things.

Justificationists always sneak in some an ad hoc, poorly specified, unstated-and-hidden-from-criticism version of CR into their thinking, which is why they are able to think at all.

This is what you were doing when saying you clarified that meant Aubreyism step 1 to include creative and critical thinking.

So what you really do is some CR, then sometimes stop and ignore some criticisms. The justificationism in the remaining steps is an excuse that hides what's going on, but contributes no value.

Some more on this at the end.

> I've articulated some rebuttals of Elliotism, and you've articulated a series of rebuttals of my rebuttals, but I'm finding them increasingly weak

"weak" is too vague to be answerable

> - I'm no longer seeing them as reaching my threshold of "meaningful" (i.e. requiring a new rebuttal).

This is too vague to be answerable. What's the threshold, and which arguments don't meet it?

> Rather, they seem only to reveal confusion on your part, such as elidin the difference between resolving a conflict of ideas and resolving a conflict of personalities, or ignoring what one knows

What who knows? I have not been ignoring things I know, so I'm unclear on what you're trying to get at.

> about the time it typically takes to generate a rebuttal when there is one out there to be generated. I've mentioned these problems with Elliotism and I'm not satisfied with your replies. Does that mean I should consider the discussion to be over? Not according to Elliotism, because in your view you are still coming up with abundantly meaningful rebiuttals of my rebuttals, i.e. we're nowhere near a win/win. But according to Aubreyism, I probably should, soon anyway, because I've given you a fair chance to come up with rebuttals that I find to be meaningful and you've tried and failed.

I don't know, specifically, what you're unsatisfied with.

It could help to focus on one criticism you think you're right about, and clarify what the problem is and why you think my reply doesn't solve it. Then go back and forth about it.

You mention two issues but without stating the criticism you believe is unanswered. This doesn't allow me to answer the issues.

1) You mention time for rebuttal creation. We discussed this. But at this point, I don't know what you think the problem is, how it refutes CR, and what was unsatisfactory about my explanations on the topic.

2) You mention the difference between conflicts of ideas and personalities. But I don't know what the criticism is.

Personalities consist of ideas, so in that sense there is no difference. I don't know what you would say about this – agree or disagree, and then reach what conclusion about CR.

But that's a literal answer which may be irrelevant.

I'm guessing your intended point is about the difference between getting people not to fight vs. actually making progress in a field like science. These are indeed very different. I'm aware of that and I don't know why you think it poses a problem for CR. With CR as with anything else, large breakthroughs aren't made at all times in every discussion. So what? The claim I've made is the possibility of acting only on non-refuted ideas.

> Oh dear - we seem to have a bistable situation. Elliotism is valid if evaluated according to Elliotism, but Aubreyism is valid if evaluated according to Aubreyism. How are we supposed to get out of that?

One approach is looking at real world results. What methods were behind things we all agree were substantial knowledge creation? Popper has done some analysis of examples from the history of science.

Another approach is to ask a hard epistemology question like, "How can knowledge be created?" Then see how well the different proposed epistemologies deal with it.

CR has an answer to this, but justificationism doesn't.

CR's answer is that guesses and criticism works because it's evolution, complete with replication, variation and selection. How and why evolution is able to create knowledge is well understood and has books like The Selfish Gene about it, as well as being covered well in DD's books.

Justificationism claims to be an epistemology method capable of creating knowledge. It therefore ought to either explain

1) how it's evolution

or

2) what a different way knowledge can be created is, besides evolution, and how it uses that

If you can't do this, you should reject justificationism. Not as an imperfect but pragmatic approach, but as being completely ineffective and useless at creating any knowledge.

Aubrey de Grey Discussion, 11

> I wouldn't draw a distinction there. If you don't know more criticisms, and resolved all the conflicts of ideas you know about, you're done, you resolved things. Whether you could potentially create more criticisms doesn't change that.

> OK, of everything you've said so far that is the one that I find least able to accept. Thinking of things takes time - you aren't disputing that. So, if at a given instant I have resolved all the conflicts I know about, but some of what I now think is really really new and I know I haven't tried to refute it, how on earth can I be "done"?

> As you say, you already know that you should make some effort to think critically about new ideas. So, you already have an idea that conflicts with the idea to declare yourself done immediately.
>
> If you know a reason not to do something, that's an idea that conflicts with it.

> Ah, but hang on: what do I actually know, there? You're trying to make it sound boolean by referring to "some" effort, but actually the question is how much effort.

The question is, "Have I done enough effort? Should I do more effort or stop now?" That is a boolean question.

Just mentioning a quantity in some way doesn't contradict CR.

> What I know is my past experience of how long it typically took to come up with a refutation of an idea that (before I tried refuting it) felt about as solid as the one I'm currently considering feels. That's correlation, plain and simple. I'm solely going on my hunch of how solid what I already

> know feels, or converseiy how likely it is that if I put in a certain amount of time trying to refute what I think I will succeed. So it’s quantitative. I can never claim I’m “done” until I’ve put in what I feel is enough effort that putting in a lot more would still not bring forth a rebuttal. And that estimated amount of effort again comes from extrapolation from my past experience of how fast I come up with rebuttals.
>
> To me, the above is so obvious a rebuttal

I think your rebuttal relies on CR being incompatible with dealing with any sort of quantity – a misconception I wasn't able to predict. Otherwise why would a statement of your approach be a rebuttal to CR?

It's specifically quantities of justification – of goodness of ideas – that CR is incompatible with.

> of what you said that it makes no sense that you would not have come up with it yourself in the time it took you to write the email. That’s what I meant about your answers getting increasingly weak.

We have different worldviews, and this makes it hard to predict what you'll say. It's especially hard to predict replies I consider false. I could try to preemptively answer more things, but some won't be what you would have said, and longer emails have disadvantages.

> I mean that it’s becoming easier and easier to come up with refutations of what you’re saying, and it seems to me that it’s becoming harder and harder for you to refute what I say - not that you’re finding it harder, but that the refutations you're giving are increasingly fragile. To my ear, they’re rapidly approaching the “that’s dumb, I disagree” level. And I don’t know what situation there would be that would make them sound like that to you too. You said earlier on that "It's hard to keep up meaningful criticism for long” and I said "That’s absolutely not my experience” - this is what I meant.
>
> > Justificationists always sneak in some an ad hoc, poorly specified, unstated-and-hidden-from-criticism version of CR into their thinking,

> which is why they are able to think at all.
>
> This is what you were doing when saying you clarified that meant Aubreyism step 1 to include creative and critical thinking.

Yes, absolutely. I don’t think I know what pure justificationism is, but for sure I agree (as I have since the start of our exchange) that CR is a better way to proceed than just by hunches and correlations.

Proceed by which correlations? Why those instead of other ones? How do you get from "X correlates with Y [in Z context]" to "I will decide A over B or C [in context D]"? Are any explanations involved? I don't know the specifics of your approach to correlations.

We've discussed correlations some, but our perspectives on the matter are so different that it wasn't easy to create full mutual understanding. It'll take some more discussion. More on this below.

> Thus, indeed Aubreyism is a hybrid between the two - it uses CR as a way to make decisions, but with a triage mechanism so that those decisions can be made in acceptable time. I’m fine with the idea that the triage part contributes no value in and of itself, because what it does do, instead, is allow the value from the CR part to manifest itself in real-world actions in a timely fashion.

Situation: you have 10 ideas, eliminate 5-8 with some CR tools, and run out of time to ponder.

You propose deciding between the remaining ideas with hunches. You say this is good because it's timely. You say the resulting value comes from CR + timeliness.

Why not roll dice to decide between those remaining ideas? That would be some CR, and timely. Do you think that's an equally good approach? Perhaps better because it eliminates bias.

I suspect you'll be unwilling to switch to dice. Meaning you believe the hunches have value other than timeliness. Contrary to your comments above.

What do you think?

> More generally, going back to my assertion that you do in fact make decisions in just the same way I do, I claim that this subjective, quantitative, non-value-adding evaluation of how different two conflicting positions feel in their solidity, and thus of how much effort one should put into further rebutting each of them, is an absolutely unavoidable aspect of applying CR in a timely fashion.

In my view, I explained how CR can finish in time. At this point, I don't know clearly and specifically why you think that method doesn't work, and I'm not convinced you understand the method well enough to evaluate. Last email, I pointed out that some of your comments are too vague to be answerable. You didn't elaborate on those points.

Bigger picture, let's try to get some perspective.

Epistemology is COMPLEX. Communication between different perspectives is VERY HARD.

When people have very different ideas, misunderstandings happen constantly, and patient back-and-forth is needed to correct them. Things that are obvious in one perspective will need a lot of clarification to communicate to another perspective. An especially open minded and tolerant approach is needed.

We are doing well at this. We should be pleased. We've gotten somewhere. Most people attempting similar things fail spectacularly.

You understand where I'm coming from better now, and vice versa. We know outlines of each other's positions. And we have a much more specific idea of what we do and don't agree about. We've discovered timely CR is a key issue.

People get used to talking to similar people and expect conversations to proceed rapidly. Less has to be communicated, because only differences require much communication. People often omit some details, but the other guy with many shared premises fills in the blanks similarly. People also commonly gloss over disagreements to be polite.

So people often experience communication as easy. Then when it isn't, they can get frustrated and give up in the face of misunderstandings and disagreements.

And justificationism is super popular, so epistemology conversations often seem to go smoothly. Similar to how most regular people would smoothly agree with each other that death from aging is good. Then when confronted with SENS, problems start coming up in the discussion and they don't have the skills to deal with those problems.

Talking to people who think differently is valuable. Everyone has some blind spots and other mistakes, and similar people will share some of the same weaknesses. A different person, even if worse than you, could lack some of your weaknesses. Trading ideas between people with different perspectives is valuable. It's a little like comparative advantage from economics.

But the more different someone is, the more difficult communication is. Attitudes to discussion have to be adjusted.

We should be pleased to have a significant amount of successful communication already. But the initial differences were large. There's still a lot of room to understand each other better.

I think you haven't discussed some details so far (including literally not replying to some points) – and then are reaching tentative conclusions about them without full communication. That's fine for initial communication to get your viewpoint across. It works as a kind of feeling out stage. But you shouldn't expect too much from that method.

If you want to reach agreement, or understand CR more, we'll have to get into some of those details. We now have a better framework to do that.

So if you're interested, I think we may be able to focus the discussion much more, now that we have more of an outline established. To start with:

Do you think you have an argument that makes timely CR LITERALLY IMPOSSIBLE, in general, for some category of situations? Just a yes or no is fine.

Aubrey de Grey Discussion, 12

> Just mentioning a quantity in some way doesn't contradict CR.

Fully agreed - but:

> The question is, "Have I done enough effort? Should I do more effort or stop now?" That is a boolean question.

Not really, because the answer is a continuum. If X effort is not enough and X+Y effort is enough, then maybe X+Y/2 effort is enough and maybe it isn’t. And, oh dear, one can continue that binary chop forever, which takes infinite time because each step takes finite time. I claim there’s no way to short-circuit that that uses only yes/no questions.

"Is infinite precision useful here? yes/no."

"Is one decimal enough precision for solving the problem we're trying to solve? yes/no"

You don't have to use only yes/no questions, but they play a key role. After these two above, you might use some method to figure out the answer to adequate precision. Then there'd be some more yes/no questions:

"Was that method we used a correct method to use here?"

"Is this answer we got actually the answer that method should arrive at, or did we follow the method wrong?"

"Have we now gotten one answer we're happy with and have no criticism of? Can we, therefore, proceed with it?"

> Plus, in the real world, at some point in that process one will in fact decide either that both the insufficiency of X and the sufficiency of X+Y are rebutted, or than neither of them is (which of the two depending on one's standard for what constitutes a rebuttal) - which indeed terminates the binary chop, but not usefully for a pure-CR approach.

Rebuttals are useful because they have information about the topic of interest. What to do next would depend on what the rebuttals are. Typically they provide new leads. When they don't, that is itself notable and can even be thought of as a lead, e.g. one might learn, "This is much more mysterious than I previously thought, I'll have to look for a new way to approach it and use more precision" – which is a kind of lead.

The standard of a rebuttal, locally, is: does this flaw pointed out by criticism prevent the idea from solving the problem we're trying to solve? yes/no. If no, it's not a criticism IN CONTEXT of the problem being addressed.

But the full standard is much more complicated, because you may say, "Yes that idea will solve that problem. However it will cause these other problems, so don't do it." In other words, the context being considered may be expanded.

> > Why not roll dice to decide between those remaining ideas? That would be some CR, and timely. Do you think that's an equally good approach? Perhaps better because it eliminates bias.
>
> Actually I'm fine with that (i.e., I recognise that the triage is functionally equivalent to that). In practice I only roll the dice when I think I'm sure enough that I know what the best answer is - so, roughly, I guess I would want to be rolling three dice and going one way if all of them come up six and the other way otherwise - but that's still dice-rolling.

There's a big perspective gap here.

I had in mind rolling dice with equal probability for each result.

If all you do is partial CR and have two non-refuted options, then they have equal status and should be given equal probability.

When you talk about amounts of sureness, you are introducing something that is neither CR nor dice rolling.

Also, if you felt 95% sure that X was a better approach than Y – perhaps a lot better – would you really want to roll dice and risk having to do Y, against your better judgment? That doesn't make sense to me.

Aubrey de Grey Discussion, 13

> So here’s an interesting example of what I mean. I woke up this morning and realised that there is indeed a rather strong refutation of my binary chop argument below, namely “don’t bother, just use X+Y - one doesn’t need to take exactly the minimum amount of time needed, only enough".

I object to the concept of a "strong refutation". I don't think there are degrees or quantities of refutation.

A reason "strong refutation" seems to make sense is because of something else. Often what we care about is a set of similar ideas, not a single idea. A refutation can binary refute some ideas in a set, and not others. In other words: criticisms that refute many variants of an idea along with it seem "strong".

People have some ability to guess whether it will be easy or hard to proceed by finding a workable close variant of the criticized idea. And they may not understand in detail what's going on, so it can seem like a hunch, and be referred in terms of strong or weak criticism.

But:

- Refuting more or fewer variant ideas is different than degrees of strength. Sometimes the differences matter.
- Hunches only have value when actually there's some reasonable underlying process being done that someone doesn't know how to put into words. Like this. And it's better to know what's going on so one can know when it will fail, and try to improve one's approach.
- People can only kinda estimate the prospects for CLOSE variants handling the criticism and continuing on similar to before. This gives NO indication of what may happen with less close variants.
- This stuff is pretty misleading because either you're aware of a variant idea that isn't refuted, or you aren't. And you can't actually know in advance how well variants you aren't aware of will work.

> But consider: yesterday I came up with the binary chop argument and it intuitively felt solid enough that I thought I'd spent enough time looking for refutations of it by the time I sent the email. I was wrong - and for sure I've been wrong in the same way many times in the past. But was I wrong to be sure enough of my argument to send the email? I'd say no. That's because, as I understand your definition of a refutation, I can't actually fix on a finite Y, because however large I choose Y to be I can always refute it by a pretty meaningful argument, namely by reference to past times when I (or indeed whole communities) have been wrong for a long time.

There are never any guarantees of being correct. Feeling sure is worthless, and no amount of that can make you less fallible.

We should actually basically expect all our ideas to be incorrect and one day be superseded. We're only at the BEGINNING of infinity.

The ways to deal with fallibilism are doing your best with your current knowledge (nothing special), and also specifically having methods of thinking which are designed to be very good at finding and correcting mistakes.

You've acknowledged your approach having some flaws, but think it's good enough anyway. That seems contrary to the spirit of mistake correction, which works best when every mistake found is taken very seriously.

I realize you also think something like one can't do better (so they aren't really flaws since better isn't achievable). That's a dangerous kind of claim though, and also important enough that if it was true and well understood, then there ought be books and papers explaining it to everyone's satisfaction and addressing all the counter-arguments. (But those books and papers do not exist.)

> Since we agreed some time ago that mathematical proofs are a field in which pure CR has a particularly good chance of being useful,

I consider CR equally useful in all fields. Substitute "CR" for "reason" in these sentences – which is my perspective – and you may see why.

> I direct you to the example of the “Lion and Man” problem, which was incorrectly “solved” for 25 years. It seems to me that the existence of cases where people can be wrong for a long time constitutes a very powerful refutation of the practicality of pure CR, since it means one cannot refute the argument that there is a refutation one hasn’t yet thought of. Thus, we can only answer “yes stop now” in finite time to "Have I done enough effort? Should I do more effort or stop now?” if we’ve already made a quantitative (non-boolean), and indeed subjective and arbitrary, decision as to how much risk we’re willing to take that there is such a refutation.

The possibility of being mistaken is not an argument to consider thinking about an issue indefinitely and never act. And the risk of being mistaken, and consequences, are basically always unknown.

What one needs to do is come up with a method of allocating time, with an explanation of how it works and WHY it's good, and some understanding of what it should accomplish. Then one can watch out for problems, keep an ear open for better approaches known to others, and in either case consider changes to one's method.

This is a general CR approach: do something with no proof it will work, no solidity, no feeling of confidence (or if you do feel confidence, it doesn't matter, ignore it). Instead, watch out for problems, and deal with them as they are found.

And here is a different answer: You cannot mitigate all the infinite risks that are logically possible. You can't do anything about the "anything is possible" risk, or the general risks inherent in fallibility. What you can do is think of specific categories of risks, and methods to mitigate those categories. Then because you're dealing with a known risk category, and known mitigation methods – not the infinite unknown – you can have some understanding of how big the downsides involved are and the effectiveness of time spent on mitigation. Then, considering other things you could work on, you can make resource allocation decisions.

It's only partially understood risks that can be mitigated against, and it's that partial understanding that allows judging what mitigation is worthwhile.

Aubrey de Grey Discussion, 14

> If all you do is partial CR and have two non-refuted options, then they have equal status and should be given equal probability.
>
> When you talk about amounts of sureness, you are introducing something that is neither CR nor dice rolling.

> I think you answer this with this:

> A reason "strong refutation" seems to make sense is because of something else. Often what we care about is a set of similar ideas, not a single idea. A refutation can binary refute some ideas in a set, and not others. In other words: criticisms that refute many variants of an idea along with it seem "strong".

> That's basically what I do. I agree with all you go on to say about closeness of variants etc, but I see exploration of variants (and choice of how much to explore variants) as coming down to a sequence of dice-rolls (or, well, coin-flips, since we're discussing binary choices).

I don't know what this means. I don't think you mean you judge which variants are true, individually, by coin flip.

Maybe the context is only variants you don't have a criticism of. But if several won their coin flips, but are incompatible, then what? So I'm not clear on what you're saying to do.

Also, are you saying that amount of sureness, or claims criticisms are strong or weak (you quote me explaining how what matters is which set of ideas a criticism does or doesn't refute), play no role in what you do? Only CR + randomness?

> Also, if you felt 95% sure that X was a better approach than Y – perhaps a lot better – would you really want to roll dice and risk having to do Y, against your better judgment? That doesn't make sense to me.

It makes sense if we remember that the choice I'm actually talking about is not between X and Y, but between X, Y and continuing to ruminate. If I've decided to stop ruminating because X feels sufficiently far ahead of Y in the wiseness stakes, then I could just have a policy of always going with X, but I could equally step back and acknowledge that curtailing the rumination constitutes dice-rolling by proxy and just go ahead and do the actual dice-roll so as to feel more honest about my process. I think that makes fine sense.

I think you're talking about rolling dice meaning taking risks in life - which I have no objection to. Whereas I was talking about rolling dice specifically as a decision making procedure for making choices. And that was in context of making an argument which may not be worth looking up at this point, but there you have a clarification if you want.

To try to get at one of the important issues, when and why would you assign X a higher percent (aka strength, plausibility, justification, etc) than Y or than ruminating more? Why would the percents ever be unequal? I say either you have a criticism of an option (so don't do that option), or you don't (so don't raise or lower any percents from neutral). What specifically is it that you think lets you usefully and correctly raise and lower percents for ideas in your decision making process?

I think your answer is you judge positive arguments (and criticisms) in a non-binary way by how "solid" arguments are. These solidity judgments are made arbitrarily, and combined into an overall score arbitrarily. Your defense of arbitrariness, rather than clearly explained methods, is that better isn't possible. If that's right, can you indicate specifically what aspects of CR you consider sometimes impossible, in what kinds of situations, and why it's impossible?

(Most of the time you used the word "subjective" rather than "arbitrary". If you think there's some big difference, please explain. What I see is a clear departure from objectivity, rationality and CR.)

> > The ways to deal with fallibilism
>
> Do you mean something different here than “fallibility”?

I meant fallibilism, but now that you point it out I agree "fallibility" is a clearer word choice.

> > are doing your best with your current knowledge (nothing special), and also specifically having methods of thinking which are designed to be very good at finding and correcting mistakes.
>
> Sure - and that’s what I claim I do (and also what I claim you in fact do, even though you don’t think you do).

I do claim to do this. Do you think it's somehow incompatible with CR?

I do have some different ideas than you about what it entails. E.g. I think that it never entails acting on a refuted idea (refuted in the actor's current understanding). And never entails acting on one idea over another merely because of an arbitrary feeling that that idea is better.

> > You've acknowledged your approach having some flaws, but think it's good enough anyway. That seems contrary to the spirit of mistake correction, which works best when every mistake found is taken very seriously.
>
> Oh no, not at all - my engagement in this discussion is precisely to test my belief that my approach is good enough.

Yes, but you're arguing for the acceptance of those flaws as good enough.

> I realize you also think something like one can't do better (so they aren't really flaws since better isn't achievable). That's a dangerous kind of claim though, and also important enough that if it was true and well understood, then there ought be books and papers explaining it to everyone's satisfaction and addressing all the counter-arguments. (But those books and papers do not exist.)

Not really, because hardly anyone thinks what you think. If CR were a widely-held position, there would indeed be such books and papers, but as far as I understand it CR is held only by you, Deutsch and Popper (I restrict myself, of course, to people who have written anything on the topic for public consumption), and Popper's adherence to it is not widely recognised. Am I wrong about that?

I think wrong. Popper is widely recognized as advocating CR, a term he coined. And there are other Critical Rationalists, for example:

http://www.amazon.com/Critical-Rationalism-Metaphysics-Science-Philosophy/dp/0792329600

This two volume CR book has essays by maybe 40 people.

CR is fairly well known among scientists. Example friendly familiar people include Feynman, Wheeler, Einstein, Medawar.

And there's other people like Alan Forrester (**http://conjecturesandrefutations.com**).

I in no way think that ideas should get hearings according to how many famous or academic people think they deserve hearings. But CR would pass that test.

I wonder if you're being thrown off because what I'm discussing includes some refinements to CR? If the replies to CR addressed it as Popper originally wrote it, that would be understandable.

But there are no quality criticisms of unmodified-CR (except by its advocates who wish to refine it). There's a total lack of any reasonable literature addressing

Popper's epistemology by his opponents, and meanwhile people carry on with ideas contradicting what Popper explained.

I wonder also if you're overestimating the differences between unmodified CR and what I've been explaining. They're tiny if you use the differences between CR and Justificationism as a baseline. Like how the difference between Mac and Windows is tiny compared to the difference between a computer and a lightbulb.

Even if Popper didn't exist, any known flaws to be accepted with Justificationism ought to be carefully documented by people in the field. They should write clear explanations about why they think better is impossible in those cases, and why not to do research trying for better since it's bound to fail in ways they already understand, and the precise limits for what we're stuck with, and how to mitigate the problems. I don't think anything good along these lines exists either.

> Since we agreed some time ago that mathematical proofs are a field in which pure CR has a particularly good chance of being useful,
>
> > I consider CR equally useful in all fields. Substitute "CR" for "reason" in these sentences – which is my perspective – and you may see why.
>
> Sorry, misunderstanding - what I meant was “Since mathematical proofs are a field in which I have less of a problem with a pure CR approach than with most fields, because expert consensus nearly always turns out to be rather rapidly achieved”

I don't think lack of expert consensus in a field is problematic for CR or somehow reduces the CR purity available to an individual.

There are lots of reasons expert consensus isn't reached. Because they don't use CR. Because they are more interested in promotions and reputation than truth. Because they're irrational. Because they are judging the situation with different evidence and ideas, and it's not worth the transaction costs to share everything so they can agree, since there's no pressing need for them to agree.

What's the problem for CR with consensus-low fields?

> > This is a general CR approach: do something with no proof it will work, no solidity, no feeling of confidence (or if you do feel confidence, it doesn't matter, ignore it). Instead, watch out for problems, and deal with them as they are found.
>
> Again, I can’t discern any difference in practice between that and what I already do.

Can you discern a difference between it and what most people do or say they do?

> I don’t think our disparate conclusions with regard to the merits of signing up with Alcor arise from you doing the above and me doing something different; I think they arise from our having different criteria for what constitutes a problem. And I don’t think this method allows a determination of which criterion for what constitutes a problem is correct, because each justifies itself: by your criteria, your criteria are correct, and by mine, mine are. (I mentioned this bistability before; I’ve gone back to your answer - Sept 27 - and I don’t understand why it’s an answer.)

Criteria for what is a problem are themselves ideas which can be critically discussed.

Self-justifying ideas which block criticism from all routes are a general category of idea which can be (easily) criticized. They're bad because they block critical discussion, progress, and the possibility of correction if they're mistaken.

> > And here is a different answer: You cannot mitigate all the infinite risks that are logically possible. You can't do anything about the "anything is possible" risk, or the general risks inherent in fallibility. What you can do is think of specific categories of risks, and methods to mitigate those categories. Then because you're dealing with a known risk category, and known mitigation methods – not the infinite unknown – you can have some understanding of how big the downsides involved are and the effectiveness of time spent on

> > mitigation. Then, considering other things you could work on, you can make resource allocation decisions.
>
> Same answer - I maintain that that's what I already do.

Do you maintain that what I've described is somehow not pure CR? The context I was addressing included e.g.:

> It seems to me that the existence of cases where people can be wrong for a long time constitutes a very powerful refutation of the practicality of pure CR, since it means one cannot refute the argument that there is a refutation one hasn't yet thought of.

You were presenting a criticism of CR, and when I talked about how to handle the issues, you've now said stuff along the lines of that's what you already do, indicating some agreement. Are you then withdrawing that criticism of CR? If so, do you think it's just you specifically who does CR (for this particular issue), or most people?

Or more precisely, the issue isn't really whether people do CR - everyone does. It's whether they *say* they do CR, whether they understand what they are doing, and whether they do it badly due to epistemological confusion.

Aubrey de Grey Discussion, 15

A reason "strong refutation" seems to make sense is because of something else. Often what we care about is a set of similar ideas, not a single idea. A refutation can binary refute some ideas in a set, and not others. In other words: criticisms that refute many variants of an idea along with it seem "strong".

That's basically what I do. I agree with all you go on to say about closeness of variants etc, but I see exploration of variants (and choice of how much to explore variants) as coming down to a sequence of dice-rolls (or, well, coin-flips, since we're discussing binary choices).

I don't know what this means. I don't think you mean you judge which variants are true, individually, by coin flip.

Maybe the context is only variants you don't have a criticism of. But if several won their coin flips, but are incompatible, then what? So I'm not clear on what you're saying to do.

Also, are you saying that amount of sureness, or claims criticisms are strong or weak (you quote me explaining how what matters is which set of ideas a criticism does or doesn't refute), play no role in what you do? Only CR + randomness?

The coin flips are not to decide whether a given individual idea is true or false, they are to decide between pairs of ideas. So let's say (for simplicity) that there are 2^N ideas, of which 90% are in one group of close variants and the other 10% are in a separate group of close variants. "Close", here, simply means differing only in ways I don't care about. Then I can do a knockout tournament to end up choosing a winning variant, and 90% of the time it will be in the first group. Since I don't actually care about the features that distinguish the variants within either group, only the features that distinguish the groups. I'm done. In other words, the solidity of an idea

> is measured by how many close variants it has - let’s call it the “variant density” in its neighbourhood. In practice, there will typically be numerical quantities involved in the ideas, so there will be an infinite number of close variants in each group - but if I have a sense of the variant densities in the two regions then that’s no problem, because I don’t need to do the actual tournament.

OK, I get the rough idea, though I disagree with a lot of things here.

You are proposing a complex procedure, involving some tricky math. It looks to me like the kind of thing requiring, minimum, tens of thousands of words to explain how it works. And a lot of exposure to public criticism to fix some problems and refine, even if the main points are correct.

Perhaps, with a fuller explanation, I could see why Aubreyism is correct about this and change my mind. I have some reasons not to think so, but I do try to keep an open mind about explanations I haven't read yet, and I'd be willing to look at a longer version. Does one exist?

Some sample issues where I'd want more detail include (no need to answer these now):

- Is the score the total variants anywhere, ignoring density, regions and neighborhoods? If so, why are those other things mentioned? If not, how is the score calculated?
- Why are ideas with more variants better, more likely to be true, or something like that? And what is the Aubreyism thing to say there, and how does that concept work in detail?
- The "regions" discussed are not regions of space. What are they, how are they defined, what are they made out of, how is distance defined in them, how do different regions connect together?
- The coin flipping procedure wouldn't halt. So what good is it?
- I can imagine skipping the coin flipping procedure because the probabilities will be equally distributed among the infinite ideas. But then the probabilities will all be infinitesimal. Dealing with those infinitesimals requires explanation.
- I'm guessing the approach involves grouping together infinitesimals by region. This maybe relies on there being a finite number of regions of ideas involved, which is a premise requiring discussion. It's not obvious because

we're looking at all ideas in some kind of idea-space, rather than only looking at the finite set of ideas people actually propose (as Elliotism and CR do normally do).

- When an idea has infinite variants, what infinity are we talking about? Is it in one-to-one correspondence with the integers, the reals, or what? Do all ideas with infinite variants have the same sort of infinity variants? Infinity is really tricky, and gets a lot worse when you're doing math or measurement, or trying to be precise in a way that depends on the detailed properties of infinity.
- There are other ways to get infinite variants other than by varying numerical quantities. One of these approaches uses conjunctions – modify an idea by adding "and X". Does it matter if there are non-numerical ways to get infinite variants? Do they make a difference? Perhaps they are important to understanding the number and density of variants in a region?
- Are there any cases where there's only finite variants of an idea? Does that matter?
- You can't actually have 90% or 10% of 2^N and get a whole number. This won't harm the main ideas, but I think it's important to fix detail errors in one's epistemology (which I think you agree with: it's why you specified 2^N ideas, instead saying even or leaving unspecified).
- Do ideas actually have different numbers of variants? Both for total number, and density. How does one know? How does one figure out total variant count, and density, for a particular idea?
- How is the distance between two ideas determined? Or whatever is used for judging density.
- What counts as a variant? In common discussion, we can make do with a loose idea of this. If I start with an idea and then think about a way to change it, that's a variant. This is especially fine when nothing much depends on what is a variant of what. But for measuring solidity, using a method which depends on what is a variant of what, we'll need a more precise meaning. One reason is that some variant construction methods will eventually construct ALL ideas, so everything will be regarded as a variant of everything else. (Example method: take ideas in English, vary by adding, removing or modifying one letter.) Addressing issues like this requires discussion.
- Where does criticism factor into things?
- What happens with ideas which we don't know about? Do we just proceed as if none of those exist, or is anything done about them?

- Does one check his work to make sure he calculated his solidity measurements right? If so, for how long?
- Is this procedure truth-seeking? Why or why not? Does it create knowledge? If so, how? Is it somehow equivalent to evolution, or not?
- Why do people have disagreements? Is it exclusively because some people don't know how to measure idea solidity like this, because of calculation errors, and because of different ideas about what they care about?
- One problem about closeness in terms of what people care about is circularity. Because this method is itself supposed to help people decide things like what to care about.
- How does this fit with DD's arguments for ideas that are harder to vary? Your approach seems to favor ideas that are easier to vary, resulting in more variants.
- I suspect there may be lots of variants of "a wizard did it". Is that a good idea? Am I counting its variants wrong? I admit I'm not really counting but just sorta wildly guessing because I don't think you or I know how to count variants.

That is only an offhand sampling of questions and issues. I could add more. And then create new lists questioning some of the answers as they were provided. Regarding what it takes to persuade me, this gives some indication of what kind of level of detail and completeness it takes. (Actually a lot of precision is lost in communication.)

Does this assessment of the situation make sense to you? That you're proposing a complex answer to a major epistemology problem, and there's dozens of questions about it that I'd want answers to. Note: not necessarily freshly written answers from you personally, if there is anything written by you or others at any time.

Do you think you know answers to every issue I listed? And if so, what do you think is the best way for me to learn those full answers? (Note: If for some answers you know where to look them up as needed, instead of always saving them in memory, that's fine.)

Or perhaps you'll explain to me there's a way to live with a bunch of unanswered questions – and a reason to want to. Or maybe something else I haven't thought of.

> > To try to get at one of the important issues, when and why would you assign X a higher percent (aka strength, plausibility, justification, etc) than Y or than ruminating more? Why would the percents ever be unequal? I say either you have a criticism of an option (so don't do that option), or you don't (so don't raise or lower any percents from neutral). What specifically is it that you think lets you usefully and correctly raise and lower percents for ideas in your decision making process?
>
> > I think your answer is you judge positive arguments (and criticisms) in a non-binary way by how "solid" arguments are. These solidity judgments are made arbitrarily, and combined into an overall score arbitrarily.
>
> I think my clarification above of the role of “variant density” as a measure of solidity answers this, but let me know if it doesn’t.

I agree with linking issues. Measuring solidity (aka support aka justification) is a key issue that other things depend on.

It's also a good example issue for the discussion below about how I might be persuaded. If I was persuaded of a working measure of solidity, I'd have a great deal to reconsider.

> Sure - and that’s what I claim I do (and also what I claim you in fact do, even though you don’t think you do).

> > I do claim to do this [quoted below]. Do you think it's somehow incompatible with CR?

> On reflection, and especially given your further points below, I’d prefer to stick with Aubreyism and Elliotism rather than justificationism and CR, because I’m new to this field and inadequately clear as to precisely how the latter terms are defined, and because I think the positions we’re debating between are our own rather than other people’s.

OK, switching terminology.

Do you think

> doing your best with your current knowledge (nothing special), and also specifically having methods of thinking which are designed to be very good at finding and correcting mistakes.

is incompatible with Elliotism? How?

> OK - as above, let’s forget unmodified CR and also unmodified justificationism. I think we’ve established that my approach is not unmodified justificationism, but instead it is (something like) CR triaged by justificationism. I’m still getting the impression that your stated approach, whether or not it’s reeeeally close to CR, is unable to make decisions adequately rapidly for real life, and thus is not what you actually do in real life.

I don't know what to do with that impression.

Do you believe you have a reason Elliotism could not be timely in theory no matter what? Or only a reason Elliotism is not timely today because it's not developed enough and the current approach is flawed, but one day there might be a breakthrough insight so that it can be timely?

I think the timeliness thing is a second key issue. If I was persuaded Elliotism isn't or can't be timely, I'd have a lot to reconsider. But I'm pretty unclear on the specifics of your counter-arguments regarding timeliness.

> > What's the problem for CR with consensus-low fields?
>
> Speed of decision-making. The faster CR leads to consensus in a given field, the less it needs to be triaged.

OK, I have a rough idea of what you mean. I don't think this is important to our main disagreements.

> This is a general CR approach: do something with no proof it will work, no solidity, no feeling of confidence (or if you do feel confidence, it doesn't matter, ignore it). Instead, watch out for problems, and deal with them as they are found.

> Again, I can’t discern any difference in practice between that and what I already do.

> Can you discern a difference between it and what most people do or say they do?

> Oh, sure - I think most people are a good deal more content than me to hold pairs of views that they recognise to be mutually incompatible.

What I was talking about above was an innocent-until-proven-guilty approach to ideas, which is found in both CR and Elliotism (without requiring infallible proof). You indicated agreement, but now bring up the issue of holding contradictory ideas, which I consider a different issue. I am unclear on whether you misunderstood what I was saying, consider these part of the same issue, or what.

Regarding holding contradictory ideas, do you have a clear limit? If I were to adopt Aubreyism, how would I decide which mutually incompatible views to keep or change? If the answer involves degrees of contentness, how do I calculate them?

Part of the Elliotism answer to this issue involves context. Whether ideas relevantly contradict each other is context dependent. Out of context contradictions aren't important. The important thing is to deal with relevant contradictions in one's current context. Put another way: deal with contradictions relevant to choices one makes.

Consider the contradicting ideas of quantum mechanics and general relativity. In

a typical dinner-choosing context, neither of those ideas offers a meal suggestion. They both say essentially "no comment" in this context, which doesn't contradict. They aren't taking different sides in the dinner arbitration. I can get pizza for dinner without coming into conflict with either of those ideas.

On the other hand if there was a contradiction in context – basically meaning they are on disagreeing sides in an arbitration – then I'd address that with a win/win solution. Without such a solution, I could only proceed in a win/lose way and the loser would be part of me. And the loser would be chosen arbitrarily or irrationally (because if it weren't, then what was done would be a rational solution and we're back to win/win).

Understanding of context is one of the things which allows Elliotism to be timely. (A refutation of my understanding of context is another thing which would lead to me reconsidering a ton.)

If I were to change my mind and live by Aubreyism, I would require a detailed understanding of how to handle context under Aubreyism (for meals, contradictions, and everything else).

> I don’t think our disparate conclusions with regard to the merits of signing up with Alcor arise from you doing the above and me doing something different; I think they arise from our having different criteria for what constitutes a problem. And I don’t think this method allows a determination of which criterion for what constitutes a problem is correct, because each justifies itself: by your criteria, your criteria are correct, and by mine, mine are. (I mentioned this bistability before; I’ve gone back to your answer - Sept 27 - and I don’t understand why it’s an answer.)

> Criteria for what is a problem are themselves ideas which can be critically discussed.
>
> Self-justifying ideas which block criticism from all routes are a general category of idea which can be (easily) criticized. They're bad because they block critical discussion, progress, and the possibility of correction if they're mistaken.

> OK then: what theoretical sequence of events would conclude with you changing your mind about how you think decisions should be made, in favour of my view?

Starting at the end, I'd have to understand Aubreyism to my satisfaction, think it was right, think Elliotism and (unmodified) CR were both wrong. The exact details are hard to specify in advance because in the sequence of events I would change my mind about what criteria to use when deciding what ideas to favor. So I would not think Aubreyism has no known criticism, rather I'd understand and use Aubreyism's own criteria. And similarly I wouldn't be rejecting Elliotism or CR for having one outstanding criticism (taking into account context), but rather because of some reasons I learned from Aubreyism.

For that matter, I might not have to understand Aubreyism to my satisfaction. Maybe it'd teach me how to adopt ideas without understanding them to my current criteria of satisfaction. It could offer different criteria of satisfaction, but it could also offer a different approach.

So, disclaimer: the below discussion of persuasion contains Elliotist ideas. But if Elliotism is false, then I guess persuasion works some other way, which I don't know and can't speak to.

Starting more at the beginning, my ideas about Elliotism are broadly integrated into my thinking (meaning connected to other ideas). An example area where they are particularly tightly integrated is parenting and education. For ease of reference, my views are called TCS (Taking Children Seriously).

So I'd have to find out things like, if I rejected Elliotism, what views am I to adopt about parenting and education? Is Aubreyism somehow fully compatible with TCS (I don't think so)? Even if it was, I'd have to find out things like how to argue TCS in new ways using Aubreyism instead of Elliotism, there'd be changes.

To give you a sense of the integration, TCS has many essays which explicitly discuss Popper, (unmodified) CR, and Elliotism. A large part of the way TCS was created was applying CR ideas to parenting and education. And also, some TCS concepts played a significant role in creating Elliotism. In addition to TCS

learning things from CR, CR can learn from TCS, resulting in a lot of the unmodified-CR/Elliotism differences.

If I'm to change my views on Elliotism and also on TCS, I'll also have to find out why the new views are moral, not immoral (or learn a new approach to morality). I'll have to find out why thousands of written TCS arguments are mistaken, and how far the mistakes go. (Small change in perspective and way of arguing basically saves all the old conclusions? Old conclusions have to be thrown out and recreated with Aubreyism? Somewhere in between?)

And when I try to change my thinking about TCS, I'll run into that fact that it's integrated with many other ideas, so will they have to change to? And they connect to yet more ideas.

So there's this tangled web of ideas. And this is just one area of integration, Elliotism and TCS. Elliotism is also integrated with my politics. And with my opinions of philosophy books. And with my approach to social life. All this could require reevaluation in light of changes to my epistemology.

How can something like this be approached?

It takes a lot of work (which I have willingness to do). One of the general facts of persuasion is, the person being persuaded has to do the large majority of the work. I'd have to persuade myself, with hints and help from you. That is the only way. You cannot make me change my mind, or do most of the work for me.

Though, again, this is an Elliotist view which might not be applicable if you refuted Elliotism. Maybe you can tell me a different way.

(Tangentially, you may note here some incompatibilities with this perspective and how school teachers approach education.)

Another consequence of this integration is that if you persuaded me I was wrong about politics, that could pose a problem for Elliotism. I'd have to figure out where the mistakes were and their full consequences, and that process might involve rejecting Elliotism. If I decide a political idea is false, and there's a chain of ideas from it to an Elliotism idea (which there is), then I'll have to find a mistake in that chain or else rethink part of Elliotism (which is itself linked with the rest of Elliotism and more, posing similar problems). So it could be possible to change my mind about Elliotism without ever discussing it.

Integration of ideas is stabilizing in some ways. If you say I'm wrong about X, I may know a dozen implications of X which I want to figure out how to deal with. This can make it more challenging to provide a satisfactory new view. But integration is also destabilizing because if I do change my mind about X, the implications spread more easily. Persuasion about one point can cause a chain reaction. Especially if I don't block off that chain reaction with a bunch of rationalizations, irrational evasions, refusals to think about implications of ideas, willful disconnections of ideas into more isolated pieces to prevent chain reaction, and so on.

The consequences of a refutation aren't predictable in advance. Maybe it turns out that idea was more isolated than you thought – or less. Maybe you can find mistaken connections near it, maybe not. Until you work out new non-refuted positions, you don't know if it will be a tiny fix or require a whole new philosophy.

Getting back to your question: The sequence of events to change my mind would be large, and largely outside of your control. The majority of it would be outside your view, even if I tried hard to share the process. My integrity would be required.

Ayn Rand says you can't "force a mind". Persuasion has to be voluntary. It's why the person to be persuaded must actively want to learn, and take initiative in the process, not be passive.

However, you could play a critically important role. If you told me one idea (e.g. how to measure solidity), and I worked out the rest from there, you would have had a major role.

More normally, I'd work out a bit from that idea, then ask you a question or argue a point, get your answer, work out a bit more, and so on. And some of your answers would refer me to books and webpages, rather than be written fresh.

It hasn't gone like this so far because I'm experiencing the epistemology discussion as you saying things I've already considered. And frequently already had several debates about. Not exactly identical ideas, but similar in the relevant ways so my previous analysis still applies. Rather than needing to rethink

something, I've been using ideas I already know and making minor adjustments to fit the details of our conversation.

I'm also using the discussion to work on ongoing projects like trying to understand Elliotism more clearly, invent better ways to explain it, and better understand where and why people misunderstand it or disagree. I also have more tangential projects like trying to write better.

It's also being used by others who want to understand Elliotism better. People write comments and use things you or I said as a jumping off point for discussions. If you wanted, you could read those discussions and comments.

Those people are also relevant to the issue of a sequence of events in which I'd be persuaded of Aubreyism. If you managed to inspire any doubts about Elliotism, or raise any problems I didn't think I had an answer to, I would raise those issues with others and see what they said. So, via me (both writing and forwarding things), you'd have to end up persuading those people of Aubreyism too. And on the other hand, they could play a big role in persuading me of Aubreyism if they understood one of your correct points before me, and then translated it to my current way of thinking well. (The Aubreyism issue could also create a split and failure to agree, but I wouldn't expect it and I see no signs of that so far.)

I also want to differentiate between full persuasion and superficial persuasion. Sometimes people are persuaded about X pretty easily. But they haven't changed their mind about anything else, so now X contradicts a bunch of their other ideas. A common result is the persuasion doesn't last. Whereas if one is persuaded about X and then makes changes to other ideas until X is compatible with all their thinking, and there's various connections, that'd be a more full kind of persuasion that does a better job of lasting.

One reason superficial persuasion seems to work and last, sometimes, is because of selective attention. People will use idea X if and only if dealing with one particular topic, and not think about other stuff. Then for other topics, they only think about other stuff and not X. So the contradictions between their other ideas and X don't get noticed, because they only think about one or the other at a time.

This further speaks to the complexity and difficulty of rational persuasion.

Getting back to a sequence of events, I don't know a specific one in detail or I'd be persuaded now. What I know is more like the categories of events that would matter and what sorts of things have to happen. (The sequencing, to a substantial extent, is flexible. Like I could learn an epistemology idea and adjust my politics, or vice versa, the sequence can go either way. At least that's the Elliotism view.)

Trying to be more specific, here's an example. You say something I don't have an answer to. It could be about measuring solidity, but it could be about pretty much any of my views I've been explaining because I take them all seriously and they're all integrated. I investigate. I find problems with several of my related ideas. I also consider some related ideas which I don't see any problem with, so I ask you about the issue. My first question is whether you think those ideas are false and I'm missing it, or you think I'm mistaken that they are related.

Trying to fix some of these problems, I run into more problems. Some of them I don't see, but you tell them to me. I start arguing some Aubreyism ideas to others who agree with Elliotism, and learn Aubreyism well enough to win those arguments (although I have to relay back to you a few of their anti-Aubreyism arguments which I'm unable to answer myself. But the more more I do that, the more I pick up on how things work myself, eventually reaching full autonomy regarding Aubreyism). Others then help me with the task of reconciling various things with Aubreyism, such as the material in Popper's books. We do things like decide some parts can be rescued and figuring out how. Other parts have to be rejected, and we work through the implications of that and figure out where and why those implications stop. To do this well involves things like rereading books while keeping in mind some Aubreyism arguments and watching out for contradictions, and thus seeing the book material in a new way compared to prior readings with a different perspective. And it involves going back through thousands of things I and others wrote and using new Aubreyism knowledge to find errors, retract things, write new things about new positions, etc. The more Aubreyism has general principles, the better this will work – so I can find patterns in what has to change instead of dealing with individual cases.

OK, there's a story. Want to tell me a story where you change your mind?

> I don’t think anyone does CR, and I also don’t think anyone does the slightly modified CR that you think you do. I think people do a triaged version of CR, and some people do the triaging better than others.

I acknowledge that's your position.

Aubrey de Grey Discussion, 16

The other parts so far are all my emails including quotes from Aubrey de Grey. For this part, I'm posting his email. That's because I didn't quote everything when replying. Outlined quotes are older.

> A reason "strong refutation" seems to make sense is because of something else. Often what we care about is a set of similar ideas, not a single idea. A refutation can binary refute some ideas in a set, and not others. In other words: criticisms that refute many variants of an idea along with it seem "strong".

That's basically what I do. I agree with all you go on to say about closeness of variants etc, but I see exploration of variants (and choice of how much to explore variants) as coming down to a sequence of dice-rolls (or, well, coin-flips, since we're discussing binary choices).

> I don't know what this means. I don't think you mean you judge which variants are true, individually, by coin flip.
>
> Maybe the context is only variants you don't have a criticism of. But if several won their coin flips, but are incompatible, then what? So I'm not clear on what you're saying to do.
>
> Also, are you saying that amount of sureness, or claims criticisms are strong or weak (you quote me explaining how what matters is which set of ideas a criticism does or doesn't refute), play no role in what you do? Only CR + randomness?

The coin flips are not to decide whether a given individual idea is true or false, they are to decide between pairs of ideas. So let's say (for simplicity) that there are 2^N ideas, of which 90% are in one group of close variants and the other 10% are in a separate group of

> close variants. “Close”, here, simply means differing only in ways I don’t care about. Then I can do a knockout tournament to end up choosing a winning variant, and 90% of the time it will be in the first group. Since I don’t actually care about the features that distinguish the variants within either group, only the features that distinguish the groups. I’m done. In other words, the solidity of an idea is measured by how many close variants it has - let’s call it the “variant density” in its neighbourhood. In practice, there will typically be numerical quantities involved in the ideas, so there will be an infinite number of close variants in each group - but if I have a sense of the variant densities in the two regions then that’s no problem, because I don’t need to do the actual tournament.

> OK, I get the rough idea, though I disagree with a lot of things here.
>
> You are proposing a complex procedure, involving some tricky math. It looks to me like the kind of thing requiring, minimum, tens of thousands of words to explain how it works. And a lot of exposure to public criticism to fix some problems and refine, even if the main points are correct.

Not really, because the actual execution of the procedure is hugely condensed. It’s just the same as when mathematicians come up with a proof: they know that the only reason the proof is sound is because it can be reduced to set theory, but they also know that in Principia Mathematica it took a couple of hundred pages to prove that 1+1=2, so they are happy not to actually do the reduction.

> Perhaps, with a fuller explanation, I could see why Aubreyism is correct about this and change my mind. I have some reasons not to think so, but I do try to keep an open mind about explanations I haven't read yet, and I'd be willing to look at a longer version. Does one exist?

No. Sorry :-)

> Some sample issues where I'd want more detail include (no need to answer these now):

I will anyway, because all but the last two are easy (I think).

> - Is the score the total variants anywhere, ignoring density, regions and neighborhoods? If so, why are those other things mentioned? If not, how is the score calculated?

No, it’s the total number of “close” variants, defined as I did before, i.e. variants that differ only in ways that one doesn’t care about.

> - Why are ideas with more variants better, more likely to be true, or something like that? And what is the Aubreyism thing to say there, and how does that concept work in detail?

Because they have historically turned out to be. Occam’s Razor, basically.

> - The "regions" discussed are not regions of space. What are they, how are they defined, what are they made out of, how is distance defined in them, how do different regions connect together?

See above - different ideas differ in multiple ways, some of which one cares about and some of which one doesn’t, so they fall into equivalence classes, and the larger classes win.

> - The coin flipping procedure wouldn't halt. So what good is it?

I’m not with you. Why wouldn’t it halt? It’s just a knockout tournalemt starting with 2^n players. Ah, are you talking about the infinite case? There, as I say, one indeed doesn’t do the flipping, one uses the densities. A way to estimate the densities would be just to sample 100 ideas that are in one of the two competing groups and see how many are in which group.

> - I can imagine skipping the coin flipping procedure because the probabilities will be equally distributed among the infinite ideas. But then the probabilities will all be infinitesimal. Dealing with those infinitesimals requires explanation.

I think I've covered that above. Yes?

> - I'm guessing the approach involves grouping together infinitesimals by region. This maybe relies on there being a finite number of regions of ideas involved, which is a premise requiring discussion. It's not obvious because we're looking at all ideas in some kind of idea-space, rather than only looking at the finite set of ideas people actually propose (as Elliotism and CR do normally do).

I think this is all compatible with the above, since only the number of equivalence classes of ideas needs to be finite, not the number of ideas.

> - When an idea has infinite variants, what infinity are we talking about? Is it in one-to-one correspondence with the integers, the reals, or what? Do all ideas with infinite variants have the same sort of infinity variants? Infinity is really tricky, and gets a lot worse when you're doing math or measurement, or trying to be precise in a way that depends on the detailed properties of infinity.

I don't think this matters for the sampling procedure I described above.

> - There are other ways to get infinite variants other than by varying numerical quantities. One of these approaches uses conjunctions – modify an idea by adding "and X". Does it matter if there are non-numerical ways to get infinite variants? Do they make a difference? Perhaps they are important to understanding the number and density of variants in a region?

I don't think this breaks the sampling procedure either.

> - Are there any cases where there's only finite variants of an idea? Does that matter?

Not sure, and not as far as I can see.

> - You can't actually have 90% or 10% of 2^N and get a whole number. This won't harm the main ideas, but I think it's important to fix detail errors in one's epistemology (which I think you agree with: it's why you specified 2^N ideas, instead saying even or leaving unspecified).

Fair enough! - sample 128 ideas instead of 100.

> - Do ideas actually have different numbers of variants? Both for total number, and density. How does one know? How does one figure out total variant count, and density, for a particular idea?

Let me know if you think the sampling procedure doesn’t do that.

> - How is the distance between two ideas determined? Or whatever is used for judging density.

See above.

> - What counts as a variant? In common discussion, we can make do with a loose idea of this. If I start with an idea and then think about a way to change it, that's a variant. This is especially fine when nothing much depends on what is a variant of what. But for measuring solidity, using a method which depends on what is a variant of what, we'll need a more precise meaning. One reason is that some variant construction methods will eventually construct ALL ideas, so everything will be regarded as a variant of everything else. (Example method: take ideas in English, vary by adding, removing or modifying one letter.) Addressing issues like this requires discussion.

Again, I think my definitions and procedure cover this.

> - Where does criticism factor into things?

It elucidates whether two ideas differ in ways one cares about. Changing one's mind about that results in changing which equivalence class the ideas fall into.

> - What happens with ideas which we don't know about? Do we just proceed as if none of those exist, or is anything done about them?

I think that's part of the CR part of Aubreyism, rather than the triage part, i.e. one does it in the same way whether one is using Aubreyism or Elliotism.

> - Does one check his work to make sure he calculated his solidity measurements right? If so, for how long?

Ditto.

> - Is this procedure truth-seeking? Why or why not? Does it create knowledge? If so, how? Is it somehow equivalent to evolution, or not?

No it isn't/doesn't/isn't - it is the triage layer that terminates a CR effort. The CR part is what is truth-seeking and creates knowledge.

> - Why do people have disagreements? Is it exclusively because some people don't know how to measure idea solidity like this, because of calculation errors, and because of different ideas about what they care about?

All those things, sure, but probably other things too -same as for CR.

> - One problem about closeness in terms of what people care about is circularity. Because this method is itself supposed to help people decide things like what to care about.

I don’t see that that implies circularity. Recursiveness, sure, but that’s OK, isn’t it?

> - How does this fit with DD's arguments for ideas that are harder to vary? Your approach seems to favor ideas that are easier to vary, resulting in more variants.

Ah, good point. I don’t adequately recall his argument, though. Can you summarise it?

> - I suspect there may be lots of variants of "a wizard did it". Is that a good idea? Am I counting its variants wrong? I admit I'm not really counting but just sorta wildly guessing because I don't think you or I know how to count variants.

Is that, basically, DD’s "harder to vary” argument?

> That is only an offhand sampling of questions and issues. I could add more. And then create new lists questioning some of the answers as they were provided. Regarding what it takes to persuade me, this gives some indication of what kind of level of detail and completeness it takes. (Actually a lot of precision is lost in communication.)

Right.

> Does this assessment of the situation make sense to you? That you're proposing a complex answer to a major epistemology problem, and there's dozens of questions about it that I'd want answers to. Note: not necessarily freshly written answers from you personally, if there is anything written by you or others at any time.

Understood; yes it does.

> Do you think you know answers to every issue I listed? And if so, what do you think is the best way for me to learn those full answers? (Note: If for some answers you know where to look them up as needed, instead of always saving them in memory, that's fine.)
>
> Or perhaps you'll explain to me there's a way to live with a bunch of unanswered questions – and a reason to want to.

I think that's exactly what I'm doing - Aubreyism is precisely that.

> Or maybe something else I haven't thought of.

> To try to get at one of the important issues, when and why would you assign X a higher percent (aka strength, plausibility, justification, etc) than Y or than ruminating more? Why would the percents ever be unequal? I say either you have a criticism of an option (so don't do that option), or you don't (so don't raise or lower any percents from neutral). What specifically is it that you think lets you usefully and correctly raise and lower percents for ideas in your decision making process?
>
> I think your answer is you judge positive arguments (and criticisms) in a non-binary way by how "solid" arguments are. These solidity judgments are made arbitrarily, and combined into an overall score arbitrarily.

> I think my clarification above of the role of "variant density" as a measure of solidity answers this, but let me know if it doesn't.

> I agree with linking issues. Measuring solidity (aka support aka justification) is a key issue that other things depend on.

> It's also a good example issue for the discussion below about how I might be persuaded. If I was persuaded of a working measure of solidity, I'd have a great deal to reconsider.

OK - but then the question is whether yout current view permits you to change your mind about this (or indeed about anything big).

> Sure - and that's what I claim I do (and also what I claim you in fact do, even though you don't think you do).

> I do claim to do this [quoted below]. Do you think it's somehow incompatible with CR?

> On reflection, and especially given your further points below, I'd prefer to stick with Aubreyism and Elliotism rather than justificationism and CR, because I'm new to this field and inadequately clear as to precisely how the latter terms are defined, and because I think the positions we're debating between are our own rather than other people's.

> OK, switching terminology.
>
> Do you think
>
> > doing your best with your current knowledge (nothing special), and also specifically having methods of thinking which are designed to be very good at finding and correcting mistakes.
>
> is incompatible with Elliotism? How?

I think the first part is imcompatible, yes; Elliotism does not deliver doing one's best with current knowledge, because it overly favours excessive rumination.

> OK - as above, let’s forget unmodified CR and also unmodified justificationism. I think we’ve established that my approach is not unmodified justificationism, but instead it is (something like) CR triaged by justificationism. I’m still getting the impression that your stated approach, whether or not it’s reeeeally close to CR, is unable to make decisions adequately rapidly for real life, and thus is not what you actually do in real life.

I don't know what to do with that impression.

Do you believe you have a reason Elliotism could not be timely in theory no matter what? Or only a reason Elliotism is not timely today because it's not developed enough and the current approach is flawed, but one day there might be a breakthrough insight so that it can be timely?

I can’t really answer the first question, because I can’t identify the set of all possible variants of current Elliotism that you would still recognise as Elliotism. For the second question, yes, that’s what I think, and moreover I think the breakthrough in question is simply to add a triage step, which would turn it into Aubreyism.

I think the timeliness thing is a second key issue. If I was persuaded Elliotism isn't or can't be timely, I'd have a lot to reconsider. But I'm pretty unclear on the specifics of your counter-arguments regarding timeliness.

What's the problem for CR with consensus-low fields?

> Speed of decision-making. The faster CR leads to consensus in a given field, the less it needs to be triaged.

OK, I have a rough idea of what you mean. I don't think this is important to our main disagreements.

I agree.

> This is a general CR approach: do something with no proof it will work, no solidity, no feeling of confidence (or if you do feel confidence, it doesn't matter, ignore it). Instead, watch out for problems, and deal with them as they are found.

> Again, I can't discern any difference in practice between that and what I already do.

> Can you discern a difference between it and what most people do or say they do?

> Oh, sure - I think most people are a good deal more content than me to hold pairs of views that they recognise to be mutually incompatible.

> What I was talking about above was an innocent-until-proven-guilty approach to ideas, which is found in both CR and Elliotism (without requiring infallible proof). You indicated agreement, but now bring up the issue of holding contradictory ideas, which I consider a different issue. I am unclear on whether you misunderstood what I was saying, consider these part of the same issue, or what.

I think holding contradictory ideas is the same issue - it's equivalent to not watching out for problems.

> Regarding holding contradictory ideas, do you have a clear limit? If I were to adopt Aubreyism, how would I decide which mutually incompatible views to keep or change? If the answer involves degrees of contentness, how do I calculate them?

Sampling to estimate variant density, followed by deciding based on coin-flips. No it doesn’t involve degrees of contentness.

> Part of the Elliotism answer to this issue involves context. Whether ideas relevantly contradict each other is context dependent. Out of context contradictions aren't important. The important thing is to deal with relevant contradictions in one's current context. Put another way: deal with contradictions relevant to choices one makes.
>
> Consider the contradicting ideas of quantum mechanics and general relativity. In a typical dinner-choosing context, neither of those ideas offers a meal suggestion. They both say essentially "no comment" in this context, which doesn't contradict. They aren't taking different sides in the dinner arbitration. I can get pizza for dinner without coming into conflict with either of those ideas.
>
> On the other hand if there was a contradiction in context – basically meaning they are on disagreeing sides in an arbitration – then I'd address that with a win/win solution. Without such a solution, I could only proceed in a win/lose way and the loser would be part of me. And the loser would be chosen arbitrarily or irrationally (because if it weren't, then what was done would be a rational solution and we're back to win/win).
>
> Understanding of context is one of the things which allows Elliotism to be timely. (A refutation of my understanding of context is another thing which would lead to me reconsidering a ton.)

I think we agree on context. In the language of variants and equivalence classes and sampling and coin flips, the introduction of an out-of-context issue simply doubles the number of variants in each equivalence clas, so it doesn’t affect the decision-making outcome (nor the time it takes to make the decision).

> If I were to change my mind and live by Aubreyism, I would require a detailed understanding of how to handle context under Aubreyism (for meals, contradictions, and everything else).

Let me know if the above suffices.

> I don’t think our disparate conclusions with regard to the merits of signing up with Alcor arise from you doing the above and me doing something different; I think they arise from our having different criteria for what constitutes a problem. And I don’t think this method allows a determination of which criterion for what constitutes a problem is correct, because each justifies itself: by your criteria, your criteria are correct, and by mine, mine are. (I mentioned this bistability before; I’ve gone back to your answer - Sept 27 - and I don’t understand why it’s an answer.)

Criteria for what is a problem are themselves ideas which can be critically discussed.

Self-justifying ideas which block criticism from all routes are a general category of idea which can be (easily) criticized. They're bad because they block critical discussion, progress, and the possibility of correction if they're mistaken.

> OK then: what theoretical sequence of events would conclude with you changing your mind about how you think decisions should be made, in favour of my view?

Starting at the end, I'd have to understand Aubreyism to my satisfaction, think it was right, think Elliotism and (unmodified) CR were both wrong. The exact details are hard to specify in advance because in the sequence of events I would change my mind about what criteria to use when deciding what ideas to favor. So I would not think Aubreyism has no known criticism, rather I'd understand and use Aubreyism's own criteria. And similarly I wouldn't be rejecting Elliotism or CR for having one outstanding criticism (taking into account context), but rather because of some reasons I learned from Aubreyism.

For that matter, I might not have to understand Aubreyism to my satisfaction. Maybe it'd teach me how to adopt ideas without

> understanding them to my current criteria of satisfaction. It could offer different criteria of satisfaction, but it could also offer a different approach.
>
> So, disclaimer: the below discussion of persuasion contains Elliotist ideas. But if Elliotism is false, then I guess persuasion works some other way, which I don't know and can't speak to.

Right - we're back to bistability.

I know, I have a better idea. I think you mentioned some time ago that before you encountered DD you thought differently about all this. Is that correct? If so, perhaps it will help if you relate the sequence of events that led you to change your mind. Since that will be a sequence of events that actually occurred, rather than a story about a hypothetical sequence, I think I'll find it more useful.

Cheers, Aubrey

Aubrey de Grey Discussion, 17

> - Why are ideas with more variants better, more likely to be true, or something like that? And what is the Aubreyism thing to say there, and how does that concept work in detail?

Because they have historically turned out to be. Occam’s Razor, basically.

How do you know what happened historically? How does that tell you what will work in a particular case now?

What you wrote is a typical inductivist statement. The idea is there are multiple observations of history supporting the conclusion (that ideas with more variants turn out to be better). Then add an inductive principle like "the future is likely to resemble the past". Meanwhile no explanation is given for why this conclusion makes sense. Is induction what you mean?

Also that isn't Occam's Razor, which is about favoring simpler ideas. More variants isn't simpler. At least I don't think so. Simpler is only defined vaguely, which does allow arbitrary conclusions. (There have been some attempts to make Occam's Razor precise, which most people aren't familiar with, and which don't work.)

> - The coin flipping procedure wouldn't halt. So what good is it?

I’m not with you. Why wouldn’t it halt? It’s just a knockout tournalemt starting with 2^n players. Ah, are you talking about the infinite case? There, as I say, one indeed doesn’t do the flipping, one uses the densities. A way to estimate the densities would be just to sample 100 ideas that are in one of the two competing groups and see how many are in which group.

Yes I meant the infinite case. By sample do you mean a random sample? In the infinite case, how do you get a random sample or otherwise make the sample fair?

Also, could you provide an example of using your method?

> > Or perhaps you'll explain to me there's a way to live with a bunch of unanswered questions – and a reason to want to.
>
> I think that’s exactly what I’m doing - Aubreyism is precisely that.

But you just attempted to give answers to many questions, rather than tell me why those questions didn't need answers.

> Do you think
>
> > doing your best with your current knowledge (nothing special), and also specifically having methods of thinking which are designed to be very good at finding and correcting mistakes.
>
> is incompatible with Elliotism? How?

> I think the first part is imcompatible, yes; Elliotism does not deliver doing one’s best with current knowledge, because it overly favours excessive rumination.

Excessive rumination is something you – but not me – think is a consequence of Elliotism. A consequence of what specific things, for what reason, I'm unclear on. Tell me.

I wrote about how the amount of time (and other resources) used on an arbitration is tailored to the amount of time one thinks should be used. I'm not clear on what you objected to. My guess is you didn't understand, which I would have expected to take more clarifying questions.

> OK - as above, let’s forget unmodified CR and also unmodified justificationism. I think we’ve established that my approach is not unmodified justificationism, but instead it is (something like) CR triaged by justificationism. I’m still getting the impression that your stated approach, whether or not it’s reeeeally close to CR, is unable to make decisions adequately rapidly for real life, and thus is not what you actually do in real life.

> I don't know what to do with that impression.
>
> Do you believe you have a reason Elliotism could not be timely in theory no matter what? Or only a reason Elliotism is not timely today because it's not developed enough and the current approach is flawed, but one day there might be a breakthrough insight so that it can be timely?

> I can’t really answer the first question, because I can’t identify the set of all possible variants of current Elliotism that you would still recognise as Elliotism. For the second question, yes, that’s what I think, and moreover I think the breakthrough in question is simply to add a triage step, which would turn it into Aubreyism.

Why do you think Elliotism itself is lacking, rather than the lacking being in your incomplete understanding of Elliotism?

> > Part of the Elliotism answer to this issue involves context. Whether ideas relevantly contradict each other is context dependent. Out of context contradictions aren't important. The important thing is to deal with relevant contradictions in one's current context. Put another way: deal with contradictions relevant to choices one makes.
> >
> > Consider the contradicting ideas of quantum mechanics and general relativity. In a typical dinner-choosing context, neither of those ideas offers a meal suggestion. They both say essentially "no comment" in this context, which doesn't contradict. They aren't taking different

> > sides in the dinner arbitration. I can get pizza for dinner without coming into conflict with either of those ideas.
> >
> > On the other hand if there was a contradiction in context – basically meaning they are on disagreeing sides in an arbitration – then I'd address that with a win/win solution. Without such a solution, I could only proceed in a win/lose way and the loser would be part of me. And the loser would be chosen arbitrarily or irrationally (because if it weren't, then what was done would be a rational solution and we're back to win/win).
> >
> > Understanding of context is one of the things which allows Elliotism to be timely. (A refutation of my understanding of context is another thing which would lead to me reconsidering a ton.)
>
> I think we agree on context. In the language of variants and equivalence classes and sampling and coin flips, the introduction of an out-of-context issue simply doubles the number of variants in each equivalence clas, so it doesn't affect the decision-making outcome (nor the time it takes to make the decision).

What about the win/win vs win/lose issue?

> I don't think our disparate conclusions with regard to the merits of signing up with Alcor arise from you doing the above and me doing something different; I think they arise from our having different criteria for what constitutes a problem. And I don't think this method allows a determination of which criterion for what constitutes a problem is correct, because each justifies itself: by your criteria, your criteria are correct, and by mine, mine are. (I mentioned this bistability before; I've gone back to your answer - Sept 27 - and I don't understand why it's an answer.)

> Criteria for what is a problem are themselves ideas which can be critically discussed.
>
> Self-justifying ideas which block criticism from all routes are a general category of idea which can be (easily) criticized. They're bad because they

> block critical discussion, progress, and the possibility of correction if they're mistaken.

> OK then: what theoretical sequence of events would conclude with you changing your mind about how you think decisions should be made, in favour of my view?

> Starting at the end, I'd have to understand Aubreyism to my satisfaction, think it was right, think Elliotism and (unmodified) CR were both wrong. The exact details are hard to specify in advance because in the sequence of events I would change my mind about what criteria to use when deciding what ideas to favor. So I would not think Aubreyism has no known criticism, rather I'd understand and use Aubreyism's own criteria. And similarly I wouldn't be rejecting Elliotism or CR for having one outstanding criticism (taking into account context), but rather because of some reasons I learned from Aubreyism.
>
> For that matter, I might not have to understand Aubreyism to my satisfaction. Maybe it'd teach me how to adopt ideas without understanding them to my current criteria of satisfaction. It could offer different criteria of satisfaction, but it could also offer a different approach.
>
> So, disclaimer: the below discussion of persuasion contains Elliotist ideas. But if Elliotism is false, then I guess persuasion works some other way, which I don't know and can't speak to.

> Right - we're back to bistability.

I don't think there's a big problem here. I already understand some things you say, and vice versa. This can be increased incrementally.

You might want to read Popper's essay "The Myth of the Framework".

You could tell me which things you considered false from what I said, and why. I don't know which are Aubreyism-compatible and which contradict Aubreyism. And you could tell me how you think persuasion should work. It takes more communication.

> I know, I have a better idea. I think you mentioned some time ago that before you encountered DD you thought differently about all this. Is that correct? If so, perhaps it will help if you relate the sequence of events that led you to change your mind. Since that will be a sequence of events that actually occurred, rather than a story about a hypothetical sequence, I think I'll find it more useful.

Correct, but there's not much to tell. DD (and others) were available for discussion. We discussed, people learned things. There was no master plan. I don't know what you're trying to find out.

The sequence of events is discussion #1, discussion #2, discussion #6,209, etc. Part of this can still be read as email archives.

Also I spent some time thinking and reading. Early on I read *The Fabric of Reality* and **http://web.archive.org/web/20030603214744/http://www.tcs.ac/Articles/index**

Aubrey de Grey Discussion, 18

> Why are ideas with more variants better, more likely to be true, or something like that? And what is the Aubreyism thing to say there, and how does that concept work in detail?

> Because they have historically turned out to be. Occam’s Razor, basically.

> How do you know what happened historically? How does that tell you what will work in a particular case now?
>
> What you wrote is a typical inductivist statement. The idea is there are multiple observations of history supporting the conclusion (that ideas with more variants turn out to be better). Then add an inductive principle like "the future is likely to resemble the past". Meanwhile no explanation is given for why this conclusion makes sense. Is induction what you mean?

> Yes it is what I mean. I agree, we have no explanation for why the future has always resembled the past, and thus no basis for the presumption that it will continue to do so. So what? - how does Elliotism depart from that? And more particularly, how do you depart from it in your everyday life?

Popper (and DD) refuted induction. How do you want to handle this? Do you want me to rewrite the content in their books? I don't think that's a good approach.

Do you think the major points you're contradicting of Popper's (and DD's) work have been refuted, by you or someone else? If not, why reject them?

My friend thinks I should copy/paste BoI passages criticizing induction and ask if you have criticism. But I think that will encourage ad hoc replies out of context. And it's hard to judge which text to include in a quote for someone else. And I don't think you want to read from books. And I haven't gotten a clear

picture of what you want to know or what would convince you, or e.g. why you think induction works. What do you think?

> Also that isn't Occam's Razor, which is about favoring simpler ideas. More variants isn't simpler. At least I don't think so. Simpler is only defined vaguely, which does allow arbitrary conclusions. (There have been some attempts to make Occam's Razor precise, which most people aren't familiar with, and which don't work.)

Ah, I see the answer now. More variants is simpler, yes, because there's a fixed set of things that can vary, each of which is either relevant or irrelevant to the decision one is trying to make. So, having more variants is the consequence of having more things that can vary be irrelevant to the decision on is trying to make - which is the same as having fewer be relevant. Which is also the same as being harder to vary in the DD sense, if I recall it correctly.

> - The coin flipping procedure wouldn't halt. So what good is it?

I'm not with you. Why wouldn't it halt? It's just a knockout tournalemt starting with 2^n players. Ah, are you talking about the infinite case? There, as I say, one indeed doesn't do the flipping, one uses the densities. A way to estimate the densities would be just to sample 100 ideas that are in one of the two competing groups and see how many are in which group.

> Yes I meant the infinite case. By sample do you mean a random sample? In the infinite case, how do you get a random sample or otherwise make the sample fair?

Yes I mean random. I don't understand your other question - why does it matter what randomisation method I use?

The random sampling you propose is impossible to do. There is no physical process that random samples from an infinite set with equal probability.

Even setting infinity aside, I don't think your proposal was to enumerate every variant on a numbered list and then do the random sample using the list. Because why sample to estimate when you already have that list? But without a list of the ideas (or equivalent), I don't know how you suggest to do the sampling, without infinity, either.

This would be easier to comment on if it was more clear what you were proposing. And I prefer not to assume people are proposing impossible nonsense, rather than asking what they mean (whereas you think Elliotism's timeliness is impossible, and prefer to claim that without specifics, over asking more about how Elliotism works). And I won't be surprised if you now say you actually meant something that's unlike what I think sampling is, or say you don't care if the sampling is unfair or arbitrary (which I tried to ask about but didn't get a direct reply to).

It seems like your position is ad hoc and you hadn't figured out in advance how it works (e.g. working out the issues with sampling), figured out what the problems in the field to be addressed are, or researched previous attempts at similar positions or alternatives (and you don't want to research them, preferring to reinvent the wheel for some reason?).

> Also, could you provide an example of using your method?
>
> I think I've answered that above, by my explanation of why seeking the alternative with more close variants is the same as Occam's razor.

I mean an example like:

We're trying to decide what to get for dinner. I propose salmon sushi or tuna sushi. You propose pizza. We get sushi with 67% odds. Is that how it's supposed to work? (Note I only know the odds here because I have a full list of the ideas.)

But wait. I don't care what God's favorite natural number is; that's irrelevant. So there's infinite sushi variants like, "Get salmon sushi, and God's favorite natural

number is 5" (vary the number).

Now what? Each idea just turned into infinite variants. Do we now say there are 2*infinity variants for sushi, and 1*infinity for pizza? And get sushi with what odds?

Should we have a sort of competition to see who can think up the most variants for their dinner choice to increase its odds? Will people who are especially clever with powersets win arguments, since they can better manufacture variants?

Or given your comments above about hard to vary, should I perhaps claim that there are fewer types of sushi than of pizza, so sushi is the better meal?

Could you adjust the example to illustrate how your approach works? I don't know how to use it.

> Or perhaps you'll explain to me there's a way to live with a bunch of unanswered questions – and a reason to want to.

> I think that’s exactly what I’m doing - Aubreyism is precisely that.

> But you just attempted to give answers to many questions, rather than tell me why those questions didn't need answers.

> Um, sure - my answers were an explanation for why a bunch of OTHER questions don’t need answers.

What are some example questions that don't need answers?

> > Excessive rumination is something you – but not me – think is a consequence of Elliotism. A consequence of what specific things, for what reason, I'm unclear on. Tell me.

> Well, for example, I think caring about what randomisation method to use (above) is excessive rumination.

I think you're dramatically underestimating the complexity of epistemology and the importance of details, and treating epistemology unlike you treat biology. In science, I think you know that details matter, like what sampling method is used in an experiment. And in general know that seemingly minor details can change the results of experiments, and can't just be ignored.

I think you see epistemology as a field where smart amateurs can quickly make stuff up that sounds about right and reasonably expect to do as well as anyone, whereas you wouldn't treat biology that way. You don't treat epistemology like a rigorous science.

This is common. Many scientists make statements straying into epistemology and other areas of philosophy (and sometimes even politics), and claim their scientific expertise still applies (and many people in the audience seem to accept this). They don't recognize field boundaries accurately, or recognize that there is a lot to learn about philosophy (or politics) that wasn't in their science education. This happens routinely.

A good example was Estep and other scientists wrote a criticism of SENS which discussed a bunch of philosophy of science (which is a sub-field of epistemology). No one writing it even claims philosophy credentials. Yet they act like they're writing within their expertise, not outside it. This was then judged by expert judges, none of whom were selected for having philosophy expertise. This is then presented as expert discussion even though there's a bunch of philosophy discussion but no philosophy experts. Look at their own summary:

http://www2.technologyreview.com/sens/docs/estepetal.pdf

> 1) SENS is based on the scientifically unsupported speculations of Aubrey de Grey, which are camouflaged by the legitimate science of others; 2) SENS bears only a superficial resemblance to science or engineering; 3) SENS and de Grey’s writings in support of it are riddled with jargon- filled misunderstandings and misrepresentations; 4) SENS’ notoriety is due almost entirely to its emotional appeal; 5) SENS is pseudoscience. We base these conclusions on our extensive training and individual and collective

> hands-on experience in the areas covered by SENS, including the engineering of biological organisms for the purpose of extending life span.

2,4,5 are primarily philosophy issues. 1 and 3 are more of a mix because they partly raise issues of whether some specific scientific SENS arguments are correct. Then after making mostly philosophy claims, they say they base their conclusions on their scientific expertise. (Note: how or whether to base conclusions is an epistemology issue too.)

Then you thought I'd have to rely on your answer to Estep to find fault with his paper, even though philosophy is my field.

Do you see what I'm talking about? My position is that philosophy is a real field, which has knowledge and literature that matter. And you won't understand it if you don't treat it that way. What do you think?

I think my interest in the sampling method is a consequence of my mathematical knowledge, not of Elliotism.

It won't have been excessive even if I'm mistaken, because if I'm mistaken (and you know better) then I'll learn something. Or do you think it would be somehow excessive to want to learn about my mistake, if I'm wrong?

I don't see how I could use Aubreyism (on purpose, consciously) without knowing how to do the sampling part. That strikes me as pretty important, and I don't understand how you expect to gloss it over. I also don't see why I should find Aubreyism appealing without having an answer to my arguments about sampling (and some other arguments too).

Regardless, if there was a reason not to question and ruminate about some category of things, I could learn that reason and then not do it. So excessive rumination would not be built into Elliotism. It wouldn't be a problem with Elliotism, only potentially a problem with my ignorance of how much to ruminate about what.

Elliotism says that "how much to ruminate about what" is a topic open to knowledge creation. How will making the topic open to critical thinking lead to the wrong answer? What should be done instead?

So I ask again: why is excessive rumination a consequence of Elliotism? Which part of Elliotism causes or requires it? (And why don't you focus more on finding out what Elliotism is, before focusing on saying it's bad?)

> > I wrote about how the amount of time (and other resources) used on an arbitration is tailored to the amount of time one thinks should be used. I'm not clear on what you objected to. My guess is you didn't understand, which I would have expected to take more clarifying questions.
>
> Maybe I don’t understand, but what you’ve seemed to be saying about that is what I’m saying is identical to what I do - triaging what you elsewhere describe as Elliotism, by reaching a point where you’re satisfied not to have answers.

I think you don't understand, and have been trying to teach me induction (among other things), and arguing with me. Rather than focusing on the sort of question-asking, misunderstanding-and-miscommunication-clearing-up, and other activities necessary to learn a complex philosophy like CR or Elliotism.

This is something I don't know how to handle well.

One difficulty is I don't know which parts of my explanations you didn't understand, and why. I've tried to find out several times but without much success. Without detailed feedback on my initial explanations, I don't know what to change (e.g. different emphasis, different details included, different questions and criticisms answered) for a second iteration to explain it in a way more personalized to your worldview. Communicating about complex topics and substantial disagreements typically requires many iterations using feedback.

I did try explaining some things multiple ways. But there are many, many possible ways to explain something. Going through a bunch semi-randomly without feedback is a bad approach.

I think there's also confusion because you don't clearly and precisely know what your position is, and modify it ad hoc during the discussion – often trying to incorporate points you think are good without realizing how they contradict

other aspects of your position (e.g incorporating DD's epistemology for hard to vary, while using Occam's razor which is contradicted by DD's epistemology). Above you say, "Ah, I see the answer now," (regarding redefining Occam's Razor after introducing it) indicating that you're working out Aubreyism as you go along and it's a moving target. This nebulous and changing nature makes Aubreyism harder to differentiate from other positions, and also serves to partially immunize it from criticism by not presenting clear targets for criticism. (And it's further immunized because you accept things like losing, arbitrariness and subjectivity – so what's left to criticize? Even induction, which Popper says is an impossible myth, becomes possible again if you're willing to count reaching arbitrary conclusions as "induction".)

By contrast, my epistemology position hasn't changed at all during this discussion, and has targets for criticism such as public writing.

Also your figure-stuff-out-as-you-go approach makes the discussion much longer than if you knew the field and your position when we started. I don't mind, but it becomes unfair when you blame the discussion length on me and complain about it. You think I ask too many questions. But I don't know what you think I should do instead. Make more assumptions about what your positions are, and criticize those?

An example is you say you use some CR. But CR is a method of dealing with issues, of reaching conclusions. So what's left to do after that? Yet you, contrary to CR, want to have CR+triage. (And this while you don't really know what CR is.) And then you advocate justificationism and induction, both of which contradict the CR you claim to be (partly) using. I don't know what to make of this without asking questions. Lots of questions, like to find out how you deal with these issues. I could phrase it more as criticism instead of questions, but questions generally work better when a position is vague or incomplete.

(Why didn't I mention all of these things earlier? Because there's so many things I could mention, I haven't had the opportunity to discuss them all.)

Perhaps I should have written more meta discussion sooner, more like I've done in this email, rather than continuing to try in various ways to get somewhere with substantive points. DD for one would say I shouldn't be writing meta discussion even now. There are a bunch of ways meta discussion is problematic. Perhaps you'll like it, but I'm not confident.

One of DD's common strategies would be to delete most of what you write every email and ask a short question about a point of disagreement, and then repeat it (maybe with minor variations, or brief comments on why something isn't an answer) for the next three emails, without explaining why it matters. Usually ends badly. Here's an example of how I could have replied to you, in full:

> On Nov 2, 2014, at 9:22 AM, Aubrey de Grey wrote:
>
> > On 28 Oct 2014, at 02:39, Elliot Temple wrote:
> >
> > > In the infinite case, how do you get a random sample or otherwise make the sample fair?
> >
> > why does it matter what randomisation method I use?
>
> Do you believe that all possible sampling methods would be acceptable?
>
> If not, then in the infinite case, how do you get a random sample or otherwise make the sample fair?

This approach controls the discussion, avoids meta discussion, and is short. If you want me to write to you in this style, I can do that. But most people don't like it. It also needs a larger number of iterations than is necessary with longer emails.

I instead (in broad strokes) tried to explain where I was coming from earlier on, and now have been trying to explain why your position is problematic, and throughout I've tried to answer your questions and individual points you raise. Meanwhile you do things like ask what would persuade me, but don't answer what would persuade you. And you talk about how Aubreyism works while not asking many questions about how Elliotism works. And you make claims (e.g. about Elliotism having a timeliness flaw) and I respond by asking you questions to try to find out why you think that, so I can answer, so then you talk about your ideas more instead of finding out how Elliotism works.

I let this happen. I see it happening, see problems with it, but don't know how to fix it. I'm more willing than you to act like a child/learner/student, ask questions

and not control discussion. And I have more patience. I don't think this discussion flow is optimal, but I don't know what to do about it. I don't know how to get someone to ask more questions and try to learn more. Nor do I know how to explain something to someone, so that they understand it, without adequate feedback and questions regarding my initial explanation, to give me some indication of where to go with iteration 2 (and 3 and 4). When the feedback is vague or non-specific, or sometimes there is none, then what is one to say next? Tough problem.

Big picture, one can't force a mind, and one can't provide the initiative or impetus for someone to learn something. People make their own choices. I think it's mostly out of my hands. Sometimes I try to explain to people what methods they'll have to use if they want to learn more (e.g. ask more questions), but it usually goes badly, e.g. b/c they say "Well maybe you should learn more" (I'm already trying to, very hard, and they aren't, and they're trying to lie about this reality) or they just don't do it and don't tell me what went wrong.

> > Why do you think Elliotism itself is lacking, rather than the lacking being in your incomplete understanding of Elliotism?
>
> I could equally ask "Why do you think Elliotism itself is not lacking, rather than the lacking being in your incomplete understanding of Elliotism?".

I'm open to public debate about this, with all comers. I've been taking every reasonable step I can figure out to find out about these things, while also being open to any suggestions from anyone about other steps to take.

Additionally, I have studied the field. In addition to reading things like Popper, I've also read about other approaches. And have sought out discussion with many people who disagree. I've made an extensive effort to find out what alternative views there are, and what's good about them, and what criticisms they have relevant to CR and Elliotism.

This includes asking people if they know anything to look into more, anyone worth talking to, etc. And looking at all those leads. It also includes work by

others besides myself. There has been a collaborative effort to find any knowledge contrary to Popper.

E.g. an Australian Popperian looked over the philosophy books being taught in the Australian universities to check for anything good. He later checked over 200 university philosophy curriculums, primarily from the US, using their websites. Looking for new ideas, new leads, material not already refuted by Popper, material that may answer one of Popper's arguments, anything unexpected, and so on. (Nothing good was found.)

This is not to say Elliotism is perfect, but I've made an extensive effort to find and address flaws, and continue to make such an effort. If there are any flaws, no one knows them, or they're keeping the information to themselves. (Or in your case, we can consider the matter pending, but so far you haven't presented any new challenge to CR or Elliotism.)

What I've found is there are a lot of CR and Elliotism arguments which no one has refutations of. But e.g. there are no unanswered inductivist arguments.

A more parallel question to ask me is why I think induction is lacking, rather than the lacking being with my understanding of induction. The reason is because I've made every effort to find out about induction and how it works and what defenses of it exist for the criticisms I have.

Induction could be better than I know – but in that case it's also better than any inductivist knows, too. It's better in some unimagined way which no one knows about. (Or maybe some hermit knows and hasn't told anyone.)

The current state of the debate – which I've made every effort to advance, and which anyone may reply to whenever they want – is that induction faces many unanswered questions and criticisms, while CR/Elliotism don't. Despite serious and responsible effort, I have been unable to find any inductivist or writing with information to the contrary.

Whereas with Elliotism, you're just initially encountering it and don't know much about it (or much about the rest of the field), so I think you should have a more neutral undecided view.

None of these things would be a major issue if you wanted to simply debate some points, in detail, to a conclusion. But they become major issues when you

consider giving up on the discussion, try to form an opinion without answering some of my arguments, think questioning aspects of your position is excessive rumination, don't want to read some arguments relevant to your claims (which is like a form of judging ideas by source instead of content. You treat the sources of written in a book by Popper or on a website by Elliot differently than the source of written in an email by Elliot), etc.

> Recall: my claim is that you actually perform Aubreyism, you just don’t realise it. It could be that I understand Elliotism better than you, just as it could be that you understand it better than I. Right?

Elliotism is not defined by what I actually do.

For example, if what I actually do involves any induction ever, then Elliotism is false. In that case, you'd be right about that and I'd be wrong. But that wouldn't mean you understand what Elliotism is better than me.

> How could we know? Using Aubreyism, we’d know by looking at how you and I have actually made decisions, changed our minds etc in the past, and comparing those actions with the descriptions of Aubreyism and Elliotism. Using Elliotism as you describe it, I’m not sure how we would decide.

If you could find any counter-example to Elliotism from real life, that would refute it.

By a counter-example I mean something that contradicts Elliotism, not merely something Elliotism says is unwise. If I or anyone else did something Elliotism says is impossible, Elliotism would be false.

If it turned out that I wasn't very good at doing Elliotism, but did nothing that contradicts what Elliotism claims about reality, then it could still be the case that people can and should do exclusively Elliotism.

What I (and you) personally do has little bearing on the issues of what epistemology is true.

A different way to approach these things is critical discussion focusing on what explanations and logic make sense. What should be done, and why? What's

possible to do? What plans about what to do are actually ambiguous and ill-defined?

For example, induction is a lot like saying, "Take a bunch of data points. Plot them on a graph. Now draw a curve connecting them and continue it along the paper too. Now predict that additional data points will (likely) fall on that curve." But there are infinite such curves you could draw, and induction doesn't say which one to draw. That ambiguity is a big non-empirical problem. (Some people have tried to specify which curve, but there are problems with their answers.)

Note this initial argument about induction, like all initial arguments, doesn't cover everything in full. Because I don't know which additional details are important to your thinking, and there's far too many to include them all indiscriminately. The way to get from initial statements of issues to understanding generally involves multiple rounds of clarifying questions.

> What about the win/win vs win/lose issue?

I go with arbitrary win/lose, i..e. coin flips.

Do you understand that that doesn't count as a "solution" for BoI's "problems are soluble"? By a solution DD means only a win/win solution. But you're trying to make losing and non-solutions a fundamental feature of epistemology, contrary to BoI. Do you have some criticisms of BoI? Do you think DD was mistaken not to include a chapter about how most problems will never be solved and you have to find a way to go through life that copes with losing in regard to most issues that come up?

Or instead of asking questions, should I simply state that you're contradicting BoI, have no idea what you're talking about, and ought to reread it more carefully? And add that I've seen the same misconceptions with many other beginners. And add that people who read books quietly on their own often come away with huge misunderstandings, so what you really need to do is join the Fallible Ideas discussion group and post public critical analysis as you go along (not non-specific doubts after finishing the book). It's important to discuss the

parts of BoI you disagree with – using specific quotes while having the context fresh in memory – and it's important to do this with BoI's best advocates who are willing to have public discussions (they can be found on FI list, which was created by merging BoI list, TCS list, and a few others). If I was more pushy like this, would that help? I'm capable of a variety of styles and approaches, but have had difficulty soliciting information about what would actually be helpful to you, or what you want. This style involves less rumination, drawn-out discussion, etc. I'm guessing you won't appreciate it or want to refute its claims. What would you like? Tell me.

> > You might want to read Popper's essay "The Myth of the Framework".
>
> I might, but on the other hand I might consider the time taken to do so to be a case of excessive rumination.

What would it take to persuade you of Elliotism or interest you in reading about epistemology? What would convince you Aubreyism is mistaken?

For example, will the sampling issue get your attention? Or will you just say to sample arbitrarily using unstated (and thereby shielded from criticism) subjective intuition? You've already recommended doing things along those lines and don't seem to mind, so what would you mind?

> > You could tell me which things you considered false from what I said, and why. I don't know which are Aubreyism-compatible and which contradict Aubreyism. And you could tell me how you think persuasion should work. It takes more communication.
>
> Quite - maybe, excessively more.

How am I supposed to answer your objections if you don't tell them to me? Or if I'm not to answer them, what do you expect or want to happen?

> What I was asking was, can you concisely summarise a particular, concrete thing about which your mind was changed? - a specific question (ideally a yes/no question) that you answer differently now that you did before you encountered DD and his ideas. And then can you summarise (as concisely as possible) how you came to view his position as superior to yours. I'm presuming that the thing will be a thing about how to make decisions, so your answer to the second question needs to be couched in terms of the decision-making method that you favoured prior to changing your mind.

Yes/no question: Is recycling a good idea? The typical residential stuff where you sort your former-trash for pickup.

My old position: yes.

DD's position: no.

What happened? A few arguments, like pointing out the human cost of the sorting. Links to some articles discussing issues like how much energy recycling plants use and how some recycling processes are actually destroying wealth. Answers to all questions and criticisms I had about the new position (I had some at the time, but don't remember them now).

Another thing I would do is take an idea I learned and then argue it with others who don't know it. Then sometimes I'd find I could win the argument no problem. But other times I'd run into some further issue to ask DD about.

In other words: arguments and discussion. That's it. There's no magic formula. You seem to think there are lessons to be learned from my past experience and want to know what they are. But I already incorporated them into Elliotism (and into my explanation of how persuasion can happen) to the extent that I know what they are. To the extent I missed something, I will be unable to tell you that part of my experience, even if I remember it, because I don't know it's important and I can't write everything down including every event I regard as unimportant.

If you want raw data, so you can find the parts you think are important, there are archives available. But if you want summary from me, then it's going to contain what I regard as the important parts, basically discussion, answering all criticisms and questions, reading supplementary material, etc, all the stuff I've been talking about.

The story regarding epistemology is similar to above, except spread out over many questions and over years. And it involves a lot of mixing of issues, rather than going one topic at a time. E.g. discussing parenting and education, or politics. Epistemology has implications for those fields, *and vice versa*.

One thing I can add, that I think was really helpful, is reading lots of stuff DD wrote (anywhere, to me or not). That provided a good examples and showed what level of precise answering of all the issues is reasonably achievable. Though not fully at first – it takes a lot of skill not to miss 95% of what he's doing and getting right. And it takes skill to ask the right questions or otherwise find out more than his initial statement (there's always much more, though many people don't realize that). Early on, even if one isn't very good at this, one can read discussions he had with others and see what questions and counter-arguments they tried and see what happened, and see how DD always has further answers, and see what sorts of replies are productive, and so on. One can gradually get a better feel for these things and build up skill.

By an effort, people can understand each other and reality better. There's no shortcut. That's the principle, and it's my history. If you want to learn philosophy, you can do that. If you'd rather continue with ideas about how life is full of losing in arbitrary ways and induction, which are refuted in writing you'd rather skip reading, you can do that instead.

Aubrey de Grey Discussion, 19

> Hi Elliot - I’m in a busy phase right now so apologies for brevity. To me the purpose of our debate is to answer the question “Is Aubrey coming to substantively incorrect conclusions about what to do or say (such as about cryonics) as a result of using epistemologically invalid methods of reasoning?”. I’m not interested in the question “Is Aubrey’s method of reasoning epistemologically invalid?” except insofar as it can be shown that I would come to different conclusions (but in the same amount of time) if I adopted a different strategy. Similarly, I’m not interested in the question "Is Aubrey coming to incorrect conclusions about what to do or say (such as about cryonics) as a result of having incomplete information/understanding about things OTHER than what method of reasoning is best?” (which seems to be what happened to you in relation to recycling,

Sort of. If I'd had a better approach to reasoning, I could have found out about recycling sooner. If I hadn't already been learning a better method of reasoning, I might have stayed in favor of recycling after seeing those articles, and many other people have done. I think you're trying to create a distinction I disagree with, where you don't give reasoning methods credit in most of life, even though they are involved with everything.

> and was also what happened to me in relation to my career change from computer science), because such examples consist only in switching to triage at a point that turned out to be premature (I could have discovered in my teens that biologists were mostly not interested in aging, which is all I needed to know in order to decide that I should work on aging rather than AI, but I didn’t consider that possibility), not in having a triage step per se. I’m quite sure that epistemology is hard, but I’m not interested in what’s epistemologically valid unless there is some practical result for my choices.

OK I see where you're coming from better now.

> It’s the same as my attitude to the existence of God: I am agnostic, not because I’ve cogitated a lot and decided that the theist and atheist positions are too close to call, but because I know I’m already doing God’s work for reasons unrelated to my beliefs, hence it makes no difference to my life choices what my beliefs are. I’m perfectly happy to believe that induction can be robustly demonstrated to be epistemologically invalid - in fact, as I said before, I already think it seems to be - but why should I care? - you haven’t told me.

Because misunderstanding how knowledge is created (in science and more generally) blocks off ways of making progress. It makes it harder to learn anything. It slows down biology and every other field. More below.

> I’m surprised at your statement about random sampling - I mean, clearly the precision of the fairness will be finite, but equally clearly the precision can be arbitrarily good, so again I don’t see why it bothers you - but again, I also don't see why I should care that I don’t see, because you haven’t given me a practical reason to care, i.e. a reason to suspect that continuing the debate may lead to my coming to different conclusions about what to do or say in the future (about cryonics or anything else).

I don't know how you propose to do arbitrarily good sampling, or anything that isn't terrible. That isn't clear to me at all, nor to several people I asked. I think it's a show-stopper problem (one of many) demonstrating the way you actually think is nothing like your claims.

I don't know how many steps I can skip for this and still be understood. You seem bored with this issue, so let's try several. I think you're assuming you have a fair ordering, and that arbitrarily fair/accurate information occurs early in the ordering. And you decide what's a fair ordering by knowing in advance what answer you want, so the sampling is pointless.

> I’ll just answer this specific point quickly:
>
> > We're trying to decide what to get for dinner. I propose salmon sushi or tuna sushi. You propose pizza. We get sushi with 67% odds. Is that

> how it's supposed to work? (Note I only know the odds here because I have a full list of the ideas.)
>
> But wait. I don't care what God's favorite natural number is; that's irrelevant. So there's infinite sushi variants like, "Get salmon sushi, and God's favorite natural number is 5" (vary the number).
>
> Now what? Each idea just turned into infinite variants. Do we now say there are 2*infinity variants for sushi, and 1*infinity for pizza? And get sushi with what odds?

Sory for over-brevity there. What we do is we put the numbers in some order, and for each number N we double the number of variants for each of sushi and pizza by adding “God’s favourite number is N” and “God’s favourite number is not N" - so the ratio of numbers of variants always stays at 2. I can’t summon myself to care about the difference between countably and uncountably infinite classes, in case that was going to be your next question.

I think you missed some of the main issues here, e.g. that getting sushi with 67% odds is a stupid way to handle that situation. It doesn't deal with explanations or criticism (why should we get which food? does anyone mind or strongly object? stuff like that is important). And it's really really arbitrary, like I could mention two more types of sushi and now it's 80% odds? Why should the odds depend on how many I mention like that? That's a bad way of making decisions. I was trying to find out what you're actually proposing to do that'd be more reasonable.

Also sampling in the infinite case is irrelevant here because you knew you wanted a 67% result beforehand (and your way of dealing with infinity here consists of just doing something with it that gets your predetermined answer).

I do think the different classes of infinity matter, because your approach implies they matter. You're the one who wanted numbers of variants to be a major issue. That brings up issues like powersets, like it or not. I think the consequences of fixing your approach to fully resolve that issue are far reaching, e.g. no longer looking at numbers of ideas. And then trying to figure out what to do instead.

> More generally, you’re absolutely right that I’m making this up as I go along - I’m figuring out why what I do works. What do I mean by

> “works”? - I simply mean, I’ve found over the years that I rarely (though certainly not never) make decisions or form opinions that I later revise, and that as far as I can see, that’s not because I’m not open to persuasion or because I move to triage too soon, but because I have a method for forming opinions that really truly is quite good at getting them right, and in particular that it’s a good balance (pretty much as good as it can be) between reliability of the decision and time to make it.

From my perspective, you're describing methods that couldn't work. So whether you were a good thinker or a bad one, you wouldn't be describing what you actually do. This matters to the high-value possibility of critical discussion and improvement of your actual methods.

BTW here is another argument that you don't think the way you claim: What you're claiming is standard stuff, not original. But we agree you think better than most people. So wouldn't you be doing something different than them? But your statements about how you think don't capture the differences.

> Take this debate. I’ve given you ample opportunity to come up with reasons why my advocacy for signing up for cryopreservation is mistaken. Potential reasons fall into two classes: data that I didn’t have (or didn’t realise I had) that affects the case, and flaws in my reasoning methods that have resulted in my drawing incorrect conclusions from the data I did have. You’ve been focusing me on the latter, and I’ve given you extended opportunity to make your case, because you’re (a) very smart and articulate and fun to talk to and (b) aligned with someone else I greatly admire. But actually all you’ve ended up doing is being frustrated by the limited amount of time I’m willing to allocate to the debate (even though for someone as busy as me it wasn’t very limited at all). That’s not actually all you’ve done, of course - from my POV, the main thing you’ve done is reinforce my confidence that the way I make decisions works well, by failing to show me a practical case where it doesn’t.

I'm not frustrated. I like you. I'm trying to speak to important issues unemotionally.

If I were to be frustrated, it would not be by you. I talk to a lot of people. I bet you can imagine that most are much more frustrating than you are.

Suppose I were to complain that people don't want to learn to think better, don't want to contribute to philosophy, don't want to learn the philosophy that would let them go be effective in other fields, don't want to stop approximately destroying the minds of approximately all children, etc.

Would I be complaining about you? No, you'd be on the bottom of the list. You're already doing something very important, and doing it well enough to make substantial progress. For the various non-SENS issues, others ought to step up.

Further, I don't know that talking with me will help with SENS progress. On the one hand, bad philosophy has major practical consequences (more below). But on the other hand, if you see things more my way, it will give you less common ground with your donors and colleagues. One fights the war on aging with the army he has, now not later. If the general changes his worldview, but no one else does, that can cause serious problems.

Maybe you should stay away from me. Reason is destabilizing (and seductive), and maybe you – rightly – have higher priorities. While there are large practical benefits available (more below), maybe they shouldn't be your priority. People went to space and built computers while having all sorts of misconceptions. If you think current methods are enough to achieve some specific SENS goals, perhaps you're right, and perhaps it's good for someone to try it that way.

So no I'm not frustrated. I can't damn you, whatever you do. I don't know what you should do. All I can do is offer things on a voluntary basis.

The wrong way of thinking slows progress in fields. Some examples:

The social sciences keep doing inadequately controlled, explanationless, correlation studies because they don't understand the methods of making scientific progress. They're wasting their time and sharing false results.

Quantum physicists are currently strongly resisting the best explanation (many worlds). Then they either try to rationalize very bad explanations (like Copenhagen theory) or give up on explanations (i.e. shut up and calculate). This puts them in a very bad spot to improve physics explanations.

AI researchers don't understand what intelligence is or how knowledge can be created. They don't understand the jump to universality, conjectures and refutations, or the falseness of induction and justificationism. They're trying to solve the wrong problems and the field has been stuck for decades.

Philosophers mostly have terrible ideas and make no progress. And spread those bad ideas to other fields like the three examples above.

Feynman offers some examples:

http://neurotheory.columbia.edu/~ken/cargo_cult.html

> I explained to her that it was necessary first to repeat in her laboratory the experiment of the other person--to do it under condition X to see if she could also get result A, and then change to Y and see if A changed. Then she would know the the real difference was the thing she thought she had under control.
>
> She was very delighted with this new idea, and went to her professor. And his reply was, no, you cannot do that, because the experiment has already been done and you would be wasting time. This was in about 1947 or so, and it seems to have been the general policy then to not try to repeat psychological experiments, but only to change the conditions and see what happened.

Repeating experiments is wasting time? What a stupid field that isn't going to figure anything out (and indeed it hasn't). And Feynman goes on to discuss how someone figured out how to properly control rat maze running by putting in sand so they can't hear their footsteps – and that got ignored and everyone just kept doing inadequately controlled rat studies.

What about medicine or biology? I don't know the field very well but I've seen articles saying things like:

http://articles.mercola.com/sites/articles/archive/2012/07/12/drug-companies-on-scientific-fraud.aspx

> Former drug company researcher Glenn Begley looked at 53 papers in the world's top journals, and found that he and a team of scientists could NOT

> replicate 47 of the 53 published studies—all of which were considered important and valuable for the future of cancer treatments!

Stuff like this worries me that perhaps current methods are not good enough for SENS to work. But somehow despite problems like this, tons of medicine does work. Maybe it's OK, somehow. More on this below.

http://www.ahrp.org/cms/content/view/846/94/

> Many journals don’t even have retraction policies, and the ones that do publish critical notices of retraction long after the original paper appeared —without providing explicit information as to why they are being retracted.

The article has various unpleasant stats about retractions.

> It is worth noting that the results of *most negative clinical trials are never published*—neither are they disclosed anywhere, except in sponsors’ confidential files and FDA marketing submissions.

95% confidence is useless if there were 19 unpublished failures. Even one unpublished negative result matters a lot. Not publishing negative results is a huge problem.

http://www.retractionwatch.com/2014/11/03/shigeaki-kato-up-to-28-retractions-with-three-papers-cited-nearly-700-times/

> Former University of Tokyo researcher Shigeaki Kato has notched his 26th, 27th, and 28th retractions, all in Nature Cell Biology. The three papers have been cited a total of 677 times.

Note how much work is built partly on top of falsehoods. Lots more retraction info on that blog; it's not pretty.

Note that all of these examples are relevant to fighting aging, not just the medical stuff.

You never know when a physics breakthrough will have an implication for chemistry which has an implication for biology.

You never know when progress in AI could lead to uploading people into computers and making backup copies.

Better social sciences or psychology work could have led to better ways to handle the pro-aging trance or better ways to deal with people to get large donations for SENS.

So many academic papers are so bad. I've checked many myself and found huge problems with a majority of them. And there's the other problems I talked about above. And the philosophy errors I claim matter a lot.

So, how does progress happen despite all this?

How come you're making progress while misunderstanding thinking methods? Does it matter?

Here's my perspective.

Humans are much better, more awesome, powerful and rational things than commonly thought. Fallible Gods. Really spectacular. And this is why humans can still be effective despite monumental folly. Humans are so effective that even e.g. losing 99% of their effectiveness to folly (on average, with many people being counterproductive) leaves them able to make progress and even create modern civilization.

And it's a testament to the human spirit. So many people suffer immensely, grit their teeth, and go on living – and even producing – anyway. Others twist themselves up to lie to themselves that they aren't suffering while somehow not knowing they're doing this, which is hugely destructive to their minds, and yet they go on with life too.

I think it's like Ayn Rand wrote:

> "Don't be astonished, Miss Taggart," said Dr. Akston, smiling, "and don't make the mistake of thinking that these three pupils of mine are some sort of superhuman creatures. They're something much greater and more astounding than that: they're normal men—a thing the world has never seen—and their feat is that they managed to survive as such. It does take an exceptional mind and a still more exceptional integrity to remain untouched by the brain-destroying influences of the world's doctrines, the accumulated evil of centuries—to remain human, since the human is the rational."

John Galt is a normal man. That is what's possible. You fall way short of him. Philosophy misconceptions and related issues drop your effectiveness by a large factor, but you lack examples of people doing better so the problem is invisible to you. Most people are considerably worse off than you.

The world doesn't have to be the way it is. So much better is possible. BoI says the same thing in several ways, some subtle, I don't know if you would have noticed.

People do so much stuff wrong, drop their effectiveness massively, and then have low expectations about what humans can do.

It's important to understand that if you have problems, even huge ones, you won't automatically or presumably notice them. And actually you should expect to have all sorts of problems, some huge, some unnoticed – you're fallible and only at the beginning of infinity (of infinite progress). This makes it always important to work on philosophy topics like how problems are found and solved. It should be a routine part of every life to work on that kind of thing, because problems are part of life.

Here's a specific example. **The Mitochondrial Free Radical Theory of Aging**, by Aubrey de Grey, p 85:

> In gerontology, as in any field of science, the development of a hypothesis involves a perpetual oscillation between creative and analytical thinking. Advances of understanding are rarely achieved by purely deductive analysis of existing data; instead, scientists formulate tentative and

> incomplete generalisations of that data, which allow them to identify which questions are useful to ask by further observation or experiment. ...
>
> The above is, in fact, so universally accepted as a cornerstone of the scientific method that some may wonder why I have chosen to belabor it. I have three reasons.

This is all wrong. Tons of errors despite being short and – as you say – widely accepted.

Does it matter? Well, you wouldn't have written it if you didn't think it mattered.

Since your current concern is whether my claims matter, I'm going to focus on why they do, rather than arguing why they are true. So let's just assume I'm right about everything for a minute. What are the consequences of that?

One mistake in the passage is the deduction/data false dichotomy for approaches. This has big practical consequences because people look for progress in two places, both wrong. That they figure anything out anyway is a testament – as above – to how amazing humans are.

It also speaks to how much people's actual methods differ from their stated methods. People routinely do things like say they are doing induction, like you imply in the passage. Even though induction impossible and has never been used to figure anything out a single time in human history. So then what you must actually do is think a different way, get an answer, and then credit induction for the answer.

Is this harmless? No! Lots of times they try to do induction or some other wrong method and end up with no answer. There are so many times they didn't figure anything out, but could have. People get stuck on problems all the time. Not consciously or explicitly understanding how to think is a big aspect of these failures.

Knowing the right philosophy for how to think allows one to better compare what one is doing to the right way. Everyone deviates some and there's room for improvement. Most people deviate a lot, so there's tons of room for improvement.

And understanding what you're doing exposes it to criticism better. The more thinking gets done in a hidden and misunderstood way, the more it's shielded from criticism.

Understanding methods correctly also allows a much better opportunity to come up with potentially better methods and try different things out. You could improve the state of the art. Or if someone else makes a breakthrough, then if you understand what's going on then you would be in a much better position to use his innovation.

You have an idea about a pro-aging trance. It's a sort of philosophical perspective on society, far outside your scientific expertise. How are you to know if it's right? By doing all the philosophy yourself? That'd be time consuming, and you've acknowledged philosophy is a serious and substantive field and you don't have as much expertise to judge this kind of question as you could. Could you outsource the issue? Consult an expert? That's tough. How do you know who really is a philosophy expert, and who isn't, without learning the whole field yourself? Will you take Harvard or Cambridge's word for it? I really wouldn't recommend that. Many prestigious philosophers are terrible.

What if you asked me? I think whether I said you're right or wrong about the pro-aging trance, either way, you wouldn't take my word for it. That's fine. This kind of thing is really hard to outsource and trust an answer without understanding yourself. Whatever I said, I could give some abbreviated explanations and it's possible you'd understand, but also quite possible you wouldn't understand my abbreviated explanations and we'd have to discuss underlying issues like epistemology details.

And the issue isn't just whether your pro-aging trance idea is right or not. Maybe it's a pretty good start but could be improved using e.g. an understanding of anti-rational memes.

And if it's right, what should be done about it? Maybe if you read "How Does One Lead a Rational Life in an Irrational Society?" by Ayn Rand, you'd understand that better. (Though that particular essay is hard to understand for most people. It clashes with lots of their background knowledge. To understand it, they might need to study other Rand stuff, have discussions, etc. But then

when one does understand all that stuff, it matters, including in many practical ways.)

I think millions of people won't shift mindsets as abruptly as you hope. One reason is because of anti-life philosophies, which you don't address. Which I don't think you know what those are, as I mean them.

One aspect of this is that lots of people don't like their lives. They aren't happy, they aren't having a good time. Most of them won't admit this and lie about it. And it's not like they only dislike their lives, they like some parts too, it's mixed. Anyway they don't want to admit this to themselves (or others). Aging gives them an excuse, a way out, without having to face that they don't like their lives (and also without suicide, which is taboo, and it's hard for people to admit they'd rather be dead).

There's other stuff too, which could be explained much faster if you had certain philosophical background knowledge I could reference. The point for now is there's a bunch of philosophical issues here and getting them right matters to SENS. You basically say people are rationalizing not having effective anti-aging technology, and that does happen some, but there's other things going on too. Your plan as you present it is focused on addressing the doubts that anti-aging technology is ready, but not other obstacles.

Does it matter if you're right about the pro-aging trance? Well, you think so, or you wouldn't bring it up. One reason it matters is because if the pro-aging trance doesn't end, it could prevent large-scale funding and effort from materializing. And some other things besides doubts about SENS effectiveness may also need to be addressed.

For example, there's bad parenting. This does major harm to the minds of children, leaving them less able to want and enjoy life, less able to think rationally, and so on. Dealing with these problems – possibly by Taking Children Seriously, or something focused on helping adults, or a different way – may be important to SENS getting widespread acceptance and funding. It's also important to the quality of scientists available to keep working on SENS, beyond the initial stages, as each new problem at later ages is found.

Part of what the pro-aging trance idea is telling people is there's this one major issue which people are stuck on and have a coping strategy for. And you even

present this coping as like a legitimate reasonable way to deal with a tough situation. This underplays how irrational people are, which is encouraging to donors by being optimistic. As mentioned earlier, sometimes people succeed at stuff, somehow, despite big problems, so SENS stuff could conceivably work anyway. But it may be that some of the general irrationality issues with society are going to really get in the way of SENS and need more addressing.

(And people learning epistemology is a big help in dealing with those. If people understand better how they are thinking, and how they should think, that's a big step towards improving their thinking.)

Ending Aging by Aubrey de Grey:

> The most immediately obvious actions would be to lobby for more funding for rejuvenation research, and for the crucial lifting of restrictions on federal funding to embryonic stem cell research in the United States, by writing letters to your political representatives, demanding change.

The very questionable wisdom of government science is a philosophical issue with practical consequences like whether people should actually do this lobbying. Perhaps it'd help more to lobby for lower taxes and for government+science separation instead. Or maybe it'd be better to create a high quality Objectivist forum which can teach many people about the virtues of life, of science, of separating the government from science, and more.

This is an example of a philosophical issue important to SENS. Regardless of whether you're right in this case, getting philosophical issues like this correct at a higher rate is valuable to SENS.

> I've had a fairly difficult time convincing my colleagues in biogerontology of the feasibility of the various SENS components, but in general I've been successful once I've been given enough time to go through the details. When it comes to LEV, on the other hand, the reception to my proposals can best be described as blank incomprehension. This is not too surprising, in hindsight, because the LEV concept is even further distant from the sort of scientific thinking that my colleagues normally do than my other ideas

> are: it’s not only an area of science that’s distant from mainstream gerontology, it’s not even science at all in the strict sense

Here you're trying to use philosophical skills to advance SENS. You're trying to do things like understand why people are being irrational and how to deal with it. Every bit of philosophical skill could help you do this better. Elliotism contains valuable ideas addressing this kind of problem.

OK, so, big picture. The basic thing is if you know the correct thinking methods, instead of having big misconceptions about how you think, you can think better. This has absolutely huge practical consequences, like getting more right answers to SENS issues. I've gone through some real life examples. Here are some simplified explanations to try to get across how crucially important epistemology is.

Say you're working on some SENS issue and the right thinking method in that situation involves trying five different things to get an answer. You try three of them. Since you don't know the list of things to do, you don't realize you missed two. So 40% of the time you get stuck on the issue instead of solve it.

Later you come up with a bad idea and think it over and look for flaws. You find two but don't recognize them as flaws due to philosophy misconceptions. You miss another flaw because you don't try a flaw-finding method you could have. Even if you knew that method, you still might skip it because you don't understand how thinking works, how you're thinking about an issue, and when to use that method.

Meanwhile, whenever you think about stuff, you spend 50% of your time on induction, justificationism, and other dead ends. Only half your thinking time is productive. That could easily be the case. The ratio could easily be worse than that.

And you have no experiences which contradict these possibilities. How would you know what it's like to think way more effectively, or that it's possible, from your past experiences? That you've figured out some stuff tells you nothing about what kind of efficiency rate you're thinking at. Doing better than some other people also does not tell you the efficiency rate.

These problems are the kinds of things which routinely happen to people. They can easily happen without being noticed. Or if some of the negative consequences are noticed, they can be attributed to the wrong thing. That's common. Like if a person believes he does thinking by some series of false and irrelevant steps, he'll try to figure out which of those steps has the problem and try some adjustments to those steps. Whereas if he knew how he actually thought, he'd have a much better opportunity to find and fix his actual problems.

You may find these things hard to accept. The point is, they are the situation if I'm right about philosophy. So it does matter.

Aubrey de Grey Discussion, 20

> OK look, one more time. I’m all about practicalities. I’m starting from the position that I make decisions in whats really close to the optimal way,

This claim of being close to limits of progress is completely contrary to *The Beginning of Infinity*, which you (or any writing by anyone, which you endorse) haven't offered criticism of.

> when taking into account the need to limit the time to make them. The challenges you give to my position seem to me to be no more than dancing around the practicalities - arguing that other methods are better without addressing the trade-off between quality and speed, or without addressing the magnitude of the difference (how often would you come to a better view than me because of a better reasoning mething? Once every million years?).

The typical person mistakenly accepts win/lose non-solutions on a daily basis.

The magnitude of the difference is: it's such a big issue it's qualitative, not quantitative. It's a more important difference than merely 100x better. It's John Galt vs. Jim Taggart. It's reason vs. irrationality.

The idea of a quality/speed tradeoff or compromise is a misconception. And an excuse for arbitrary irrationality. It's the kind of thing that blights people's lives on a daily basis, as well as hindering scientific progress.

There do exist quality/speed tradeoffs in some sense of the term. But NOT in the sense of ever requiring acting on arbitrary ideas, win/loses, non-solutions, or known-to-be-refuted ideas. Which is what you say you do, on a regular basis. Every time you do that it's a big mistake that Elliotism would have handled differently by finding a non-refuted non-arbitrary idea in a timely manner and using that.

> When I look back at history and I see people making mistakes, I see those mistakes arising from lack of information, or from prejudice, etc - I can’t think of a single case where the mistake arose from using induction or justificationism rather than CR.

The mistakes don't arise from lack of information. Even deep space has lots of information, like *The Beginning of Infinity* discusses.

How did Louis Pasteur refute the spontaneous generation theory? He did experiments in which he looked at the conditions under which food and wine would spoil. They wouldn't spoil unless germs got in. Why didn't anybody do those experiments before? People knew food spoiled before Louis Pasteur came along. Microscopes had been around since the 17th century. So why did it take until the mid 19th century? People weren't looking for an explanation or for a solution to the relevant problems. They had methodology problems. Huge scientific opportunities are routinely passed over, for decades (or much longer) because people are bad at philosophy, bad a thinking, bad at science.

Most inductivists have had unproductive careers, never figuring out anything very important. I'm guessing you treat it as natural that most people aren't geniuses, and miss lots of stuff. You sort of expect the status quo. But what you're used to is caused by deeply irrational thinking methods. Rational methods open up unbounded human potential.

Prejudice, etc, are epistemology-methodology issues too.

> I'd very much have expected you to raise such an example by now.

I gave several such examples in my previous email, e.g. the explanationless correlation studies in the social sciences. That's a bunch of justificationists wasting their careers using justificationist methods that will never work.

You apparently didn't understand what was being said (typically both our faults, communication is hard) and didn't ask for more explanation (your fault, big methodology error that really messes up communication, discussion and learning).

> It’s true that I’m pretty unsure whether I’m elaborating a good justification for my own methods, because after all I am making it up as I go along - but conversely I still claim that there’s a good chance that my methods to indeed withstand scrutiny (again, measured in terms of practicalities), simply because I’m unaware of any substantive changes having occurred in my methods for a good few decades.

Not having learned anything major in philosophy in the last few decades is a terrible argument that your ideas are good enough and you can stop worrying about learning.

And if you aren't having many problems in practice, it could be because you're actually doing an unrefined version of Elliotism. It's not an argument that any of the philosophy you're advocating is any good.

Your stated methods don't withstand scrutiny. Early on I criticized them. You conceded they have big flaws. Then you claimed they are practical anyway, basically because you assume better isn't possible. More recently I also pointed out (for example) that the random sampling stuff doesn't work at all, a topic you dropped without ever saying a way to do it.

There is a better way to think, you aren't at the limits of progress. So I explained it, and you said it wouldn't work in a timely fashion. Why? What's the criticism of my position? You didn't understand it well enough to answer, and also didn't ask questions and give feedback to find out more about it. And we've been kind of stuck there, plus going on some tangents to discuss some other misconceptions.

You haven't understood Elliotism's way of getting timely non-refuted non-arbitrary ideas to act on because of the very thinking methodology errors you believe are harmless. That includes e.g. being unwilling to read things explaining how to do it, which really messes up your ability to learn anything complex. Then, somehow, you blame me or my ideas when you straight up refused to make the effort required to learn something like Elliotism. If you're busy, fine, but that isn't a flaw of Elliotism or a failure on my part. It'd be you choosing not to find time to learn about something important enough to make a reasonable judgment about it.

> At the bottom line: why do you think we still, after all this discussion, disagree about cryonics?

Primarily because we're both more interested in epistemology and discussed that more. And a major feature of the cryonics part of the discussion was your epistemology view (and mine to the contrary) that it'd take too long to work out the cryonics issues in the amount of detail I think is needed to correctly judge that sort of complex issue. (An amount of detail which I think you exceed in your biology thinking.)

Secondarily because you didn't answer a lot of what I said about cryonics and resisted giving arguments (which I kept asking for) either directly criticizing my position or explaining and arguing yours. This is a result of your methodology which doesn't pay enough attention to individual precise ideas and criticisms, and instead jumps from a vague understanding to an arbitrary conclusion.

(I suspect you approach biology in a different, significantly better way. But if you understood the correct thinking methodology, and what you actually did in biology, that'd enable you to compare and make valuable refinements. So philosophy still matters.)

Aubrey de Grey Discussion, 21

> Hi Elliot - thanks again - I sincerely wish I could allocate more time to this, but I'm just not seeing the value. Yes I know that until only one or two hundred years ago essentially everyone was so bad at the scientific method that progress was much slower than it could have been, but I'm not seeing that that's the case any more. If you're saying no, we're still going a lot slower than we could because we're reasoning poorly, and if you're right, then you or others who are following your methods (such as DD, presumably) should be contributing very disproportionately to scientific progress, but I'm not seeing that happening. Cryonics is part of biology, so I'm not getting why you say I approach biology in a better way than I approach cryonics, but in any event I claim I approach all aspects of biology (including cryonics) in the same way.

DD has contributed very disproportionately to scientific progress. But that's a tiny sample size. I'm not a scientist, by choice. I don't agree with your look-at-scientific-contributions method, but in any case you don't have the input data to use it. Yet somehow you think you've gotten a conclusion from it. You're making a mistake which defends other mistakes (they can pile up like that).

Your arguments in your books about topics like mitochondria are much more detailed and rigorous than what you said to me about cryonics.

Scientific progress is much slower than it could be, today. This can be seen by surveying scientific fields. I've already given you some examples like the social sciences and medical retractions. You didn't give alternative interpretations or criticisms. Now you deny it after leaving those points unanswered, without exposing your reasoning to criticism.

Let's look at one field more closely. Quantum physics is screwed up. DD has explained:

http://vimeo.com/5490979 (First 15 minutes.)

DD says progress with Everett's theory was slow over last 50 years. He speaks to the irrational philosophy of Everett-dissenting physicists. Then he proposes philosophical mistakes *by Everett people* as the thing to change to improve the field's progress. It's like, "Most quantum physicists are using irrational philosophies and wasting their careers. But even in that context, the philosophical mistakes of the pro-Everett physicists are big enough to focus on instead."

In 2012, answering in a physics context, "What would it look like that would be different to the way things are at the moment?", DD wrote:

https://groups.yahoo.com/neo/groups/Fabric-of-Reality/conversations/topics/24387

> For instance, there'd be:
>
> In theoretical physics: Work on the structure of the multiverse, its implications for the theory of probability, deeper explanations of various quantum algorithms, deeper understanding of the Heisenberg Picture....
>
> In philosophy: Work on things like personal identity, the relationship between multiple universes and multiple copies in a single universe, morality in the multiverse...
>
> In theoretical physics, experimental physics and philosophy: Cessation of work whose only interest is in the context of believing nonsensical 'interpretations'...
>
> In physics teaching: Excision of anti-rational ideologies such as positivism or shut-up-and-calculate from physics classes.

Physicists are spending a great deal of effort on the philosophical equivalent of denying dinosaurs existed (as DD explains in the video and in BoI), rather than doing productive work on issues like those above. That slows progress dramatically.

In **BoI**, in "A Physicist’s History of Bad Philosophy", DD writes:

> READER: But then why is it that only a small minority of quantum physicists agree?

> DAVID: Bad philosophy.

DD spends the chapter explaining. No one has refuted his arguments.

Here is an example, specifically, of a bad pro-Everett paper which goes wrong epistemologically (because of justificationism not CR): **http://users.ox.ac.uk/~everett/docs/Wallace%20epistemology.pdf**

If examples like this would change your mind, more could be provided. Or if detailed criticism of this paper would change your mind, that could be provided.

So when you say science isn't going slow (and philosophy issues lack big consequences), without addressing the problems with any scientific fields, I think you're mistaken. And you're doing it in such a way that, if you are mistaken, you won't find out.

Aubrey de Grey Discussion, 22

> I don't agree with your look-at-scientific-contributions method, but in any case you don't have the input data to use it. Yet somehow you think you've gotten a conclusion from it.

I guess I was too abbreviated: what I meant was that if disproportionate scientific progress were made by those with a minority view about how to reason, it wouldn’t be the minority view for long (at least not within science), and that hasn’t happened.

This claim, dealing with a field you don't want to study, brings up dozens of difficult issues which I think you don't want to discuss to resolution. I don't know what to do with this. Do you?

I'll mention a few example issues:

- If everyone thinks like this, who will try stuff in the first place? Who will be the early adopters? Is your plan to rely on people who *disagree with you* about this matter to be the ones to find, test, and then persuade you of innovations?

- The cause of success is something people disagree about, e.g. someone might attribute DD's success to him being an outlier genius, rather than to his philosophy.

- Small sample size. And many people don't know which scientists were Popperians. Take a hypothetical scientist who admires 100 scientists who were especially effective. 70 of them might be Popperians without him knowing.

- Judging which scientists actually are Popperians is difficult and requires philosophical skill to do accurately.

- You're proposing people would do something because it makes sense to do. But sometimes people are irrational and act in ways that don't make sense.

- It's a bit like asking if capitalism is so much better, why doesn't it dominate the whole world yet? There are many things that can block the uptake of good ideas other than the idea being mistaken.

Because there are many reasons things might not work out as you propose, you shouldn't rely on that way of looking at it. Instead, the only reasonable thing to do is look at the actual merits and content of CR arguments, not the unargued reactions of others. Look at the substantive ideas and arguments, not the opinions of others.

Either you personally should consider CR ideas, or (preferably since it's not your field) others should and you could read some summary work and be persuaded by that and reference it if challenged. So CR arguments get answered (or accepted), and there is a way for you to find out about new ideas (via the work you endorse, which provides targets for criticism, being refined or refuted). But you don't want to take responsibility for this, and nor do *lots* other people, and so the the march of progress is dramatically delayed.

> > Your arguments in your books about topics like mitochondria are much more detailed and rigorous than what you said to me about cryonics.
>
> Um sure, but that’s becaue I referred you to alcor.org and cryonics.org. I deny that the arguments given there are much (indeed any) less detailed and rigorous than those I give about mitochondria etc.

You didn't want to refer me to specific material, and I was unable to find material in the same league as your stuff. I wrote to you explaining problems with some material I found (I didn't find equivalent problems in your books). If I misjudged it, or they offer better material, you could tell me.

You do things like consider all the challenges SENS has to deal with to work, and address each. Where is the equivalent cryonics material?

There is a great deal of detailed scientific knowledge about mitochondria (which you carefully studied and learned). Where is the equivalent cryonics material?

> Scientific progress is much slower than it could be, today. This can be seen by surveying scientific fields. I've already given you some examples like the social sciences and medical retractions. You didn't give alternative interpretations or criticisms. Now you deny it after leaving those points unanswered, without exposing your reasoning to criticism.

Apologies again for over-brevity. Of course there are many reasons why scientific progress is much slower than it could be, but my contention is that inferiority of scientific method is not a significant one of them. Rather, the reasons are lack of funding from public sources beholden to the public (who certainly don't reason well), self-serving short-termist competition between scientists fomented by that lack of funding, egos, that sort of thing. Theer is also a big contribution from poor interpretation (for example, poor use of statistics), but again that is not because scientists don't believe statistics should be done right, it's because they find it more important to publish than to be correct.

Here you bring up complex and controversial philosophical issues, including freedom and capitalism. What do you think I should do? Try to explain a bunch of philosophy when you have one foot out the door, while previous attempts to explain other philosophy are unresolved? Ask why you're confident in your judgments of these issues even though your philosophy is under-specified and under-studied, and you've chosen not to read a lot of the material on these topics? Guess that you might not recognize your paragraph as bringing up a bunch of complex and controversial philosophical issues, and guess what your reasoning might be, and try to preemptively answer it? Tell you that your perspective here contains mistakes relevant to SENS funding, so our philosophical differences do matter? Any suggestions?

I would know how to handle these things if we were both using my preferred methods. But you use your own methods in the discussion, and I don't know how

to work with those. I don't know how issues like these are to be resolved with your discussion methods.

You deal with philosophy issues routinely, but you don't want to study it, and nor do you want to outsource that and endorse the conclusions in some specific writing. So you end up doing a mix of reinventing half of the wheel badly, plus outsourcing-by-accident to people whose names you don't even know so there's no accountability. You're accepting a bunch of ideas (e.g. induction) that you picked up somewhere and you don't know clearly who to hold accountable, which books are involved, where to look up details of their reasoning if I question it, etc. You're outsourcing philosophy thinking third-hand: some people have ideas and others decide they were successful and still others are impressed and spread the ideas through the culture to you.

> Concerning quantum physics, I am not a specialist, but my understanding is that the Copenhagen and Everett interpretations make exactly the same predictions about observable data, and thus cannot be experimentally distinguished. My question then is, who cares which is correct? The passage you quote from [topics/24387] (https://groups.yahoo.com/neo/groups/Fabric-of-Reality/conversations/topics/24387 "") seems to me to acknowledge this: it says that the only real problem with the Copenhagen model is that it's nonsensical. What exactly is wrong with "shut up and calculate" if it works?

Did you read The Beginning of Infinity? Do you or anyone else have answers to it? Do you want me to rewrite it with less editing? Quote it? Will you be pleased with a reference to it, telling you where to get answers?

I also don't think it makes sense to drop the random sampling topic (for example) and take up this new one – won't we run into the same discussion problems again on this new topic? I expect to; do you disagree?

BoI:

> Although Schrödinger's and Heisenberg's theories seemed to describe very dissimilar worlds, neither of which was easy to relate to existing conceptions of reality, it was soon discovered that, if a certain simple rule

of thumb was added to each theory, they would always make identical predictions. Moreover, these *predictions* turned out to be very successful.

With hindsight, we can state the rule of thumb like this: whenever a measurement is made, all the histories but one cease to exist. The surviving one is chosen at random, with the probability of each possible outcome being equal to the total measure of all the histories in which that outcome occurs.

At that point, disaster struck. Instead of trying to improve and integrate those two powerful but slightly flawed explanatory theories, and to explain why the rule of thumb worked, most of the theoretical-physics community retreated rapidly and with remarkable docility into instrumentalism. If the predictions work, they reasoned, why worry about the explanation? So they tried to regard quantum theory as being *nothing but* a set of rules of thumb for predicting the observed outcomes of experiments, saying nothing (else) about reality. This move is still popular today, and is known to its critics (and even to some of its proponents) as the ‘shut-up-and-calculate interpretation of quantum theory’.

This meant ignoring such awkward facts as (1) the rule of thumb was grossly inconsistent with both theories; hence it could be used only in situations where quantum effects were too small to be noticed. Those happened to include the moment of measurement (because of entanglement with the measuring instrument, and consequent decoherence, as we now know). And (2) it was not even *self*-consistent when applied to the hypothetical case of an observer performing a quantum measurement on another observer. And (3) both versions of quantum theory were clearly describing *some* sort of physical process that *brought* about the outcomes of experiments. Physicists, both through professionalism and through natural curiosity, could hardly help wondering about that process. But many of them tried not to. Most of them went on to train their students not to. This counteracted the scientific tradition of criticism in regard to quantum theory.

Let me define ‘bad philosophy’ as philosophy that is not merely false, but actively prevents the growth of other knowledge. In this case, instrumentalism was acting to prevent the explanations in Schrödinger’s and Heisenberg’s theories from being improved or elaborated or unified.

To understand what this means more, it's important to read the whole book and engage with it's ideas, e.g. by asking questions about points of confusion or disagreement, and criticizing parts you think may be mistaken, and discussing those things to resolution. Or if you don't do that, I think you should say more "I don't know"s instead of e.g. making the philosophical claims that shut up and calculate works, Aubreyism works, etc.

I think you want to neither answer the points in BoI and elsewhere (including by endorsing someone else's answer for use as your own), nor defer to them, nor be neutral. Isn't that irrational?

Aubrey de Grey Discussion, 23

> Elliot, you seem to be missing a very fundamental point here, namely: you seem to be working from the assumption that it’s my job to refute your position to your satisfaction. That is no more my job than it is yours to refute mine to my satisfaction.

If you care about reason, that requires dealing with criticism, to resolution. Reason requires criticisms must not be ignored, they have to be addressed (not by you personally. there must be answers you endorse, whoever writes them). This is crucial to reason so that you don't continue with bad ideas indefinitely even though better ones are known. It allows error correction instead of entrenching errors.

It is your right and privilege to live a different lifestyle than this. But then you wouldn't be a rational intellectual.

> If you think that Alcor’s or CI's refutations of concerns about cryonics (the ones you’ve definitely already found, because they are in their FAQs) are less compelling than mine about SENS, you’re entitled to your opinion, but my sincere opinion is that they are every bit as compelling. I put it to you that the evaluation of how compelling an argument is is an EXTREMELY subjective thing, both to you and to me, arising essentially from how immediately a refuttion of it comes to mind. So, it is hopeless to try to agree whether this or that argument is more compelling than the other argument: each of us must make his own judgement on that, and then act on that judgement in the indicated way - by seeking more information, or by accepting a particular conclusion a likely enough to be right that further investigation is not a priority.

I don't know why you're speaking to me at all when you hold the irrationalist position that reaching agreement in truth-seeking discussion is hopeless. (I also don't know why you are sufficiently satisfied with irrationalism that you are unwilling to read the books refuting it and offering a better way.)

> Again I repeat my bottom line: you have not given me the slightest reason to believe that people's failure to adhere to CR (or to Elliotism) is appreciably slowing the progress of science and technology.

I gave you examples and explanations, which you largely didn't reply to. Then you state I gave you no reason. That's unreasonable on your part.

> Maybe I can explain what kind of reason I would accept as valid evidence for that. Arguably, when quantum theory and relativity supplanted classical physics, they did so by taking seriously the incompatibility between wave-theoretic and particle-theoretic descriptions of light, and such like, which had been basically swept under the carpet for ages. My impression is that that isn't actually an item of evidence for your position, because (a) it was a long time ago, when many fewer people were any good at science; (b) it hadn't really been all that swept under the carpet - it was just that no one had come up with a resolution; and (c) even to the extent that it had been, the key point is that there was clear data that needed to BE so swept, whereas in the case of Copenhagen versus Everett (which I'm not sure is the same as Schrödinger versus Heisenberg, but I don't think that matters for present purposes) there is no such data, since both theories make the same predictions. If I'm wrong, and the lack of a widespread adoption of a CR-like method of reasoning back then seems likely to have substantially delayed the arrival of modern physics, persuade me.

I have tried to persuade you (in a way in which I could find out I'm mistaken, too), but you are taking steps to prevent persuasion. I cannot persuade you unilaterally. What you have done includes:

- Not replying to many points and questions.
- Not giving appropriate feedback on initial statements so we can iterate to the point of you understanding what I'm saying. Miscommunication and misunderstanding are to be expected and there has to be iteration of an error-correcting process for effective communication of ideas. (Communication being necessary to me persuading you.)
- Not being willing to read things, study issues, put enough effort into learning the topics.

A specific detail: I can't reasonably be expected to persuade you about the history of science first, as you propose. What needs to happen first is you understand what is a CR-like method of reasoning, so you can accurately evaluate which scientists did that and which didn't. But you don't want to read the texts explaining what is a CR-like method of reasoning, or ask the questions to understand it. You aren't finding out from existing material or from a heavy back-and-forth process adequate to cover a large topic.

> Or take another example from the past. If you're right that science is so slowed by this, how can it be so hard to identify an example (one that isn't far more parsimoniously explained by sociological considerations such as I outlined in my last email)?

Lots of the sociological considerations are explained by the philosophical issues I'm talking about. Because you don't know what CR is, you can't tell what is a consequence of CR or non-CR.

We have, for example, an educational theory. Where does short-term thinking, bias, egos, etc come from? Significantly, from bad educational practices. Education is fairly directly an epistemology issue and CR offers some better ideas about what educational techniques work or not.

Regarding statistics, yes scientists believe they should be done right, and sometimes there are time and money issues. But lots of people don't know what doing them right means. There are philosophical misconceptions about how to use statistics correctly which would be problems even with more time and money. (An example is the inductivist misconception that correlations hint at causation, which isn't a funding issue.)

The underlying problem is you don't understand where I'm coming from and what the world would look like if I'm right. That can't be settled by looking at examples. I gave you initial statements of Elliotism. The rational way to proceed is to iterate on that (you give feedback, ask questions, I reply, etc, understanding is iteratively created) in order to understand what I'm saying.

> And remember, what I really mean here is not "science" in the DD sense, i.e. the improved "understanding" (whatever that is) of nature, but technology, i.e. the practical application of science. Computers today rely

> absolutely on the fact that we no longer adhere to classical physics, but they rely not at all on the fact that most people work with Copenhagen rather than Everett. The passage you quote from BoI totally doesn't help, because it stops at "understanding", "knowledge", "explanations" etc, which in my book are simply smoke and mirrors until and unless they translate into practical consequences for technology. Not even implemented technology - technological proposals, like SENS, would be fine.

You have an anti-philsophical outlook and don't understand the perspective of DD, me, Popper, etc. If you want to understand and address such matters, there are ways you can, which we could focus on. I've tried to indicate how that can happen, e.g. with iterative discussion of how CR works. If you'd rather simply leave critics unanswered, just tell me you don't want to talk.

> I read FoR, but I don't think I ever read BoI. Perhaps part of why is that I found FoR to be fatally flawed on about page 4, as I think I mentioned earlier. DD is a great thinker, whom I hugely admire, but that doesn't mean I think all his thinking is correct or relevant to my own priorities. And you haven't given me any new motivation to read BoI.

I don't think you mentioned that. And I just searched the discussion and I'm not finding it.

If you would say your criticism of FoR, that'd be great. When people share criticisms in public, then progress can be made. I know DD wrote the book partly in hopes of receiving such criticism so human knowledge could advance. But you and many others with similar methods withhold criticism and dodge lots of discussion and then human knowledge creation is slowed.

Sharing your FoR criticism could help advance our discussion, too. It's topical and I've been trying to get direct criticisms from you. If you tell me what is unacceptable to you, then I could address it or concede. And if I address ALL issues you have with my view, that's how persuasion would happen. Since you already accepted your view has flaws, if you had NO objections you'd accept mine.

If you're right about FoR being flawed, you have an important insight that others could learn from. If you're mistaken, by sharing your criticism you would expose it to criticism and you could learn about your error from others. If you'd prefer to retreat from rational discussion instead, that is your choice.

Aubrey de Grey Discussion, 24

> Did you receive my email?

Hi - yes, I got it, but I couldn't think of anything useful to say.

If you want to stop talking, or adjust the terms of the conversation (e.g. change the one message at a time back and forth style), please say so directly because silence is ambiguous.

But see my comments below. I don't think we're at an impasse. I think what you said here was particularly productive.

> We have reached an impasse in which you insist on objecting to my failure to address various of your points, but I object to your failure to address my main point, namely that there is no objective measure of the rebuttability of a position. I am grateful for your persistence, since it has certainly helped me to gain a better understanding of the rationality IN MY OWN TERMS, i.e. the internal consistency, of my position, and in retrospect it is only because of my prior lack of that understanding that I didn't zero in sooner on that main point as the key issue. But still it is the main point. I don't see why I should take the time to read things to convince me of something that I'm already conceding for sake of argument, i.e. that Aubreyism is epistemologically inferior to Elliotism. And I also don't see why I should take the time to work harder to convince you of the value of cryonics when you haven't given me any reason to believe that your objections (i.e. your claim that Alcor's arguments are rebuttable in a sense that my arguments for SENS are not, and moreover that that sense is the correct one) are objective.
>
> Also, for a lot of the things I haven't replied to it's because I'm bemused by your wording. To take the latest case: when I've asked you for examples where science could have gone a lot faster by using CR rather than whatever else was used, and you have cited cases that I think are far more

> parsimoniously explained by sociological considerations, you've now come back with the suggestion that "Lots of the sociological considerations are explained by the philosophical issues I'm talking about". To me that's not just a questionable or wrong statement, it's a nonsensical one. My point has nothing whatever to do with the explanations for the sociological considerations - it is merely that if you accept that other issues than the CR/non-CR question (such as the weight that rationalists give to the views of irrationalists, because they want to sleep with them or whatever) slow things down, you can't argue that the CR/non-CR question slowed them down.

When I say something you think is nonsense, if you ignore that and try to continue the rest of the conversation, we're going to run into problems. I meant what I said, and it's important to my position, so please treat it seriously. By ignoring those statements, it ends up being mysterious to me why you disagree, because you aren't telling me your biggest objections! They won't go away by themselves because they are what I think, not random accidents.

In this case, there was a misunderstanding. You took "explained" to mean, "make [a situation] clear to someone by describing it in more detail". But I meant, "be the cause of". (Both of those are excerpts from a dictionary.) I consider bad epistemology the cause of the sociological problems, and CR the solution. I wasn't talking about giving abstract explanations with no purpose in reality. I'm saying philosophy is they key issue behind this sociological stuff.

I regard this sort of passing over large disagreements as a methodological error, which must affect your discussions with many people on all topics. And it's just the sort of topic CR offers better ideas about. And I think the outcome here – a misunderstanding that wouldn't have been cleared up if I didn't follow up – is pretty typical. And misunderstandings do happen all the time, whether they are getting noticed and cleared up or not.

I think the sex issue is a great example. Let's focus on that as a representative example of many sociological issues. You think CR has nothing to do with this, but I'll explain how CR has everything to do with it. It's a matter of ideas, and CR is a method of dealing with ideas (an epistemology), and such a method (an epistemology) is necessary to life, and having a better one makes all the difference.

Chart:

Epistemology -> life ideas -> behavior/choices

What does each of those mean?

1) Epistemology is the name of one's **method of dealing with ideas**. That includes evaluating ideas, deciding which ideas to accept, finding and fixing problems with ideas, integrating ideas into one's life so they are actually used (or not), and so on. This is not what you're used to from most explicit epistemologies, but it's the proper meaning, and it's what CR offers.

2) Life ideas determine one's behavior in regard to sex, and everything else. This is stuff like one's values, one's emotional makeup, one's personality, one's goals, one's preferences, and so on. In this case, we're dealing with the person's ideas about sex, courtship and integrity.

3) Behavior/choices is what you think it means. I don't have to explain this part. In this case it deals with the concrete actions taken to pursue the irrational woman.

You see the sex example as separate from epistemology. I see them as linked, one step removed. Epistemology is one's method of dealing with (life) ideas. Then some of those (life) ideas determine sexual behavior/choices.

Concretizing, let's examine some typical details. The guy thinks that sex is a very high value, especially if the woman is very pretty and has high social status. He values the sex independent of having a moral and intellectual connection with her. He's also too passive, puts her on a pedestal, and thinks he'll do better by avoiding conflict. He also thinks he can compromise reason in his social life and keep that separate from his scientific life. (Life) ideas like these cause his sexual behavior/choices. If he had different life ideas, he'd change his behavior/choices.

Where did all these bad (life) ideas come from? Mainly from his culture including parents, teachers, friends, TV, books and websites.

Now here's where CR comes in. Why did he accept these bad ideas, instead of finding or creating some better ideas? That's because of his epistemology – his method of dealing with ideas. His epistemology let him down. It's the underlying

cause of the cause of the mistaken sexual behavior/choices. A better epistemology (CR) would have given him the methods to acquire and live by better life ideas, resulting in better behavior/choices.

Concretizing with typical examples: his epistemology may have told him that checking those life ideas for errors was unnecessary because everyone knows they're how life works. Or it told him an ineffective method of checking the ideas for errors. Or it told him the errors he sees don't matter outside of trying to be clever in certain intellectual discussions. Or it told him the errors he sees can be outweighed without addressing them. Or it told him that life is full of compromise, win/lose outcomes are how reason works, and so losing in some ways isn't an error and nothing should be done about it.

If he'd used CR instead, he would have had a method that is effective at finding and dealing with errors, so he'd end up with much better life ideas. Most other epistemologies serve to intellectually disarm their victims and make it harder to resist bad life ideas (as in each of the examples in the previous paragraph). Which leads to the sociological problems that hinder science.

Everyone has an epistemology – a method of dealing with ideas. People deal with ideas on a daily basis. But most people don't know what their epistemology is, and don't use an epistemology that can be found in a book (even if they say they do).

The epistemologies in books, and taught at universities, are mostly floating abstractions, disconnected from reality. People learn them as words to say in certain conversations, but never manage to use them in daily life. CR is not like that, it offers something better.

Most people end up using a muddled epistemology, fairly accidentally, without much control over it. It's full of flaws because one's culture (especially one's parents) has bad ideas about epistemology – about the methods of dealing with ideas. And one is fallible and introduces a bunch of his own errors.

The only defense against error is error-correction – which requires good error-correcting methods of dealing with ideas (epistemology) – which is what CR is about. It's crucial to learn about what one's epistemology is, and improve it. Or else one will – lacking the methods to do better – accept bad ideas on all topics and have huge problems throughout life.

And note those problems in life include problems one isn't aware of. Thinking your life is going well doesn't mean much. The guy with the bad approach to sex typically won't regard that as a huge problem in his life, he'll see it a different way. Or if he regards it as problematic, he may be completely on the wrong track about the solution, e.g. thinking he needs to make more compromises with his integrity so he can have more success with women.

PS This is somewhat simplified. Epistemology has some direct impact, not just indirect. And I don't regard the sociological problems as the only main issue with science, I think bad ideas about how to do science (e.g. induction) matter too. But I think it's a good starting place for understanding my perspective. Philosophy dominates EVERYTHING.

Aubrey de Grey Discussion, 25

Aubrey did not reply further.

Note my **last email** to him began by saying:

> If you want to stop talking, or adjust the terms of the conversation (e.g. change the one message at a time back and forth style), please say so directly because silence is ambiguous.

He answered this with silence.

I think that's pretty unreasonable.

I don't really want to write comments on the content of the conversation. It should speak for itself. (And in a fair way with back-and-forth discussion, rather than just me talking.)

But I did want to comment briefly on attitude to discussion.

You can read some of my thoughts on this topic in my **Paths Forward** essay. I think Aubrey is blocking discussion and preventing there from being paths forward. If he's mistaken, and it's a big deal, how will he find out when he simply leaves various criticisms unresolved and unanswered? If some of the epistemology he doesn't know is true and important, how will he ever find that out while not understanding it, not asking enough questions to understand it, and not having a refutation of it (by himself or anyone else)?

I think it's very important to address rival ideas. Either personally or by outsourcing – it's fine to use someone else's writing in place of your own, as long as you take personal responsibility for its correctness, as if it was your own. If a criticism of your position is not addressed by anyone (in public writing that's exposed to public criticism, comments, question-asking, discussion, etc), then it really ought to be addressed not ignored. Aubrey neither addresses various Popperian ideas (such as the refutation of justificationism), nor does he know of any writing by anyone else which addresses it. Yet he rejects it and stops

pursuing it, without having any answer to it. This is not symmetric. The Popperian ideas I'm advocating are exposed to public criticism but are not currently refuted by anything. My ideas meet all challenges; Aubrey's don't; and Aubrey stopped discussing, leaving it like that without changing his mind.

Aubrey de Grey Vs. Smoking

Quotes from **Ending Aging: The Rejuvenation Breakthroughs That Could Reverse Human Aging in Our Lifetime** by Aubrey de Grey.

> And, slowly but surely, smoking is becoming less popular. Just like drunk driving before it, smoking is becoming socially disreputable. It's a long, hard road, though: not just because nicotine is addictive, but because youngsters continue to take up smoking despite the social stigma increasingly attached to it.

Sometimes they smoke *because of*, not despite, that social stigma. Sometimes they want to rebel against social control.

> the battle to protect youngsters from taking up smoking is one that virtually all adults, smokers or not, support.

This is a political position which is nowhere near universal. **Not everyone** thinks children should be "protect[ed]" – meaning controlled supposedly for their own good. Some people value the freedom to smoke, and the freedom of individuals (even young individuals) to choose their own fate. Some people see some value in smoking (e.g. **South Park has defended smoking**). Some people think children should be helped to become more wise, rather than protected. Maybe good advice and control over their own lives works better for children than protection. There are diverse approaches to this topic.

Similarly, **not everyone** agrees about addiction. **I don't**.

Approaching issues by saying everyone agrees is a bad approach in general. Look what would happen with SENS and aging. People would say virtually everyone disagrees with SENS, so it's bad. The same tactic could be used against most innovative new ideas, early on.

> with smoking, even though it causes some of those self-same diseases, somehow society is itself subject to an addiction that robs it of its

> rationality concerning new young addicts. We face every day the brutal disconnect between allowing cigarettes to be advertised and sold widely and seeing how much they blight and shorten the lives of those who fall under their spell.

Rather than argue with people who disagree with him, here Aubrey de Grey attacks their rationality and metaphorically accuses them of a mental illness (addiction). He then attacks free trade and free speech, as if his positions against those things are uncontroversial and need no explanation. (Saying a product is good is speech; selling it is trade. Disallowing those things is incompatible with freedom.)

People who disagree with you are not mentally ill. They have not fallen under a magic "spell". People are capable of thinking and disagreeing with you. You should expect that and speak to the issues, rather than gloss over the issues (no direct criticism of freedom was provided) and spend your time denying the other side exists. Try to find win/win solutions which address people's concerns. Persuade people instead of calling them mentally ill, irrational, or otherwise talking around their arguments.

It'd be better to approach this **like David Deutsch**: "in every human dispute there's a substantive issue at stake". Calling the other side mentally ill does not help anyone better understand the substantive issue at stake. Claiming (correctly or not) that one's position is popular, or creating a social stigma against things one disagrees with, are not truth-seeking approaches.

Are Anti-SENS Arguments Dumb?

Biogerontologists' Duty to Discuss Timescales Publicly by Aubrey de Grey:

> ... the prevalence of comments from laypeople along the lines of “Who would want to spend all that time being old?”, “Wouldn’t we get terribly bored?” or “How would we pay for all those pensions?” fills many of us with such awe at their breathtaking stupidity that any ardour to persist in a patient explanation of what success in this endeavour would actually mean is rapidly sapped. But this is not a legitimate reaction to such inanity, in my view. To put it simply, it is just not plausible that people are really that dumb. Hence, before we abandon our fellow man to his misconception, we as biogerontologists are duty bound to seek a more satisfactory basis for the persistence of these extraordinarily transparently flawed opinions.
>
> On doing so we are forced, it seems to me, to acknowledge that one very simple reason fits the facts: denial.

But in **Ending Aging: The Rejuvenation Breakthroughs That Could Reverse Human Aging in Our Lifetime** by Aubrey de Grey:

> ... the prospect of eventually being able to combat aging as well as we can currently combat most infectious diseases—essentially to eliminate aging as a cause of death, in other words—strikes terror into most people: Their immediate (and, I must point out, often high-pitched) reaction is to raise the specter of uncontrollable overpopulation, or of dictators living forever, or of only a wealthy elite benefiting, or any of a dozen other concerns.
>
> Now, I’m certainly not saying that these objections are dumb—not at all. We should indeed be considering them as dangers that we should work to preempt by appropriately careful forward planning.

Previously (2003), Aubrey de Grey said these objections are dumb, inane, and breathtakingly stupid. Later (2007), he says they certainly aren't dumb. These statements contradict. Which is it – and why?

Previously he attacked these sorts of objections, but condescendingly defended the speakers as rationalizing not arguing. Rather than address the issues, he focused on ad hominem claims about the psychology of people who disagree with him. But four years later he says the objections are reasonable concerns which should be considered and dealt with by careful planning.

> I consider it highly likely that within ten years from now, if the rather modest necessary funding is forthcoming, we will have the ability to take a mouse cohort with a three-year life expectancy, when it is already two years old, and treble its remaining life expectancy (that is, give it a total life expectancy of five years). I also consider it highly likely that the announcement of that degree of control over mouse aging will almost instantly overturn society's prevailing fatalism concerning any chance of personal benefit from real anti-aging medicine.

The objections won't all instantly melt away because they are not just meaningless emotional irrationality. It's so condescending to think there's no real objections. It's going to take patient discussions to create agreement with the many people who currently disagree (and it should not be assumed they are wrong about everything – rational discussions must be approached without assuming the conclusions in advance). It'd be better to begin that process today, rather than expect a shortcut will work.

Improved technology simply won't answer concerns about boredom, dictators or overpopulation. Nor will the objections be addressed by calling them dumb and then commenting negatively about the objectors, rather than discussing the issues to find win/win solutions. Condescendingly calling others irrational is itself an irrational way to deal with intellectual issues.

Letter to SENS

I sent the below letter to **SENS**, which is a medical research non-profit seeking to solve human aging. I like them because they have a good plan for how to do this which makes sense. Aubrey de Grey is their leader, I had a long discussion with him which you can **read here**.

SENS claims to be basically the most important thing in the world. SENS' web presence is inconsistent with this claim. SENS' web presence communicates low-prestige, low-intellectual-seriousness amateur hour. I offer criticism for several issues, partly on Aubrey's direct invitation, in hopes of helping.

Concrete Examples:

The SENS website LOOKS like a very standard generic format that doesn't stand out at all or get attention.

The SENS website has many basic web design errors such as:

- requires giving your country and even US State to sign up for newsletter. email should be the ONLY required field, period. and don't even ask for stuff like people's zip code. it's not OK to add friction to newsletter signups.

- SENS front page should be aimed at the public. that means you don't put things like "jobs" and "terms of use" there. you put all the stuff the public doesn't care about on an About page or other internal page.

- the February newsletter webpage does not link to the previous newsletter, or the archives, at the bottom.

- SENS has 3 blogs instead of 1 blog with categories. this splits up viewer attention. and since all 3 are very inactive, it just makes them look even more inactive – even with triple content in one place it'd still look bad and like SENS is inactive.

- It just plain looks like a cheap generic site in terms of layout and design. It's hard to explicitly explain why it does, but lots of people can tell because they've seen many other websites that look similar. The look of the site doesn't stand out and doesn't DIFFERENTIATE SENS. It doesn't communicate that this is something special or important.

- The images used look generic and unimpressive too. They don't stand out.

- It's not a .com site. That's bad because lots of people don't understand other TLDs besides com. (People given the website URL in person will literally do things like try to go to sens.org.com or just forget and go to sens.com. This especially applies to older people who I'm guessing are a larger part of the SENS audience. This issue is well known and makes a substantial difference.)

- The site doesn't have a bunch of awesome impressive essays (or other content) with amazing ideas. Or if it does they aren't prominent and I managed to miss them.

The SENS newsletter isn't even consistently once per month (which would be the bare minimum frequency to not look bad and have people forget about you).

The SENS newsletter looks like a normal newsletter, it doesn't stand out, it doesn't communicate SENS is SUPER FUCKING IMPORTANT.

The SENS contact form looks like a generic "we have to put up a contact form to pretend we listen to feedback" black hole. I don't know whether it is or not, but it looks that way. It looks generic and boring, and like you won't get a reply just like you don't from many other organizations. And it even adds annoying friction like making you categorize your inquiry – which is asking people, if they want to contact SENS at all, to do extra work which they aren't good at and don't want to do.

The SENS website homepage links to the SENS subreddit. This is not OK because that subreddit is very inactive (the 15th highest submission is 3 months old!). Do not send homepage visitors to a dead site, only link them places they should actually go and will be glad they went.

When you claim SENS is super duper important, but lots of the stuff you do implicitly contradicts, you destroy your own credibility and drive away most

people.

Here's an example of acting inconsistently with your **claims from Facebook**:

Jonathan Weaver That's $10,000 in 2-3 days. Nice booster.

Like · Reply · December 5, 2014 at 6:45am

SENS Foundation Jonathan Weaver That's right! We're very thankful.

Like · Reply · December 5, 2014 at 8:31am

SENS claims to need something like $100,000,000/yr for the RMR project to go full speed and save everyone's lives. 10k/2.5 days would be too little by a factor of 68 if you got it constantly all year. 10k fundraising also just looks bad for being a small amount of money, all kinds of unimportant projects get more than 10k on kickstarter in 2-3 days. By being happy with a small amount, you accept it as appropriate to SENS, and accept a status below all sorts of stuff that can raise more.

If you really think you need 100mil/yr or MILLIONS OF PEOPLE DIE (which is what even a few year delay for SENS means), then sound the alarm instead of saying you're happy with an amount of fundraising that kills millions. When you act happy with pennies, you are telling people SENS isn't really that big a deal.

You may doubt the importance of these things. Keep in mind the cultural context. People don't expect to be listened to. If SENS is any different (which I'm unclear on), you have to shout it from the rooftops before anyone will notice. You have to make the difference extremely clear.

When Joe Random has what he thinks is a good idea, he knows he'll have a hell of a time getting anyone to listen, be it a big company, a small company, a scientist, a politician, etc. It's true that the majority of Joe Randoms have bad ideas, but some have good ideas and some others could learn to have good ideas with some pointers in the right directions. If you want Joe to communicate with SENS, you have to get his attention, not blend in with every other organization that he expects to ignore him.

I posted at the **subreddit** per **Aubrey's recommendation** and got replies which said, basically:

1) Leave and email Aubrey personally (or Michael Rae or SENS) instead.

2) Leave and go to the longevity subreddit which is more active. [Note: the longevity subreddit isn't really active either.]

3) I like SENS but got discouraged from the SENS subreddit because my posts kept getting downvoted.

4) You could try posting here and hope that somehow things will work out, contrary to your reasonable expectation.

I was not impressed. And the subreddit does nothing to stand out and communicate SENS IS IMPORTANT.

I think the talk to Aubrey/Michael personally plan is problematic because they are busy. For SENS to succeed on a big scale, there needs to be division of labor rather than expecting Aubrey/Michael to do most stuff personally. It also communicates that SENS is small time and un-prestigious if it doesn't have anyone below the top people to answer questions and have discussions with the public – there should be tiers with only a few things being escalated to the top people.

I checked the SENS Facebook page that Aubrey mentioned. It, again, does nothing to stand out and communicate that SENS is something different that's really important. It's more active than the subreddit. I dislike Facebook so I'm not familiar enough with Facebook pages to say if the activity level is OK or not, but it's definitely not GREAT.

I'd like to differentiate between three different styles of promoting SENS. Three categories of how to approach this. SENS is not doing well for any of them.

Style 1) Prestige

Impress people and say how SENS is smarter than you, and works with prestigious people and has a fancy reputation, etc, etc

This is irrational and will alienate the best and smartest people, but will impress the second tier people. It could work I guess (I'm not a fan of this style and don't recommend it).

SENS does some stuff clearly in this direction, but overall isn't good at this. An **example** in this style is writing, "Extramural research at PRESTIGIOUS universities and other state-of-the-art laboratory facilities throughout the world". Which isn't even well done, it's crude and blatant. Achieving prestige works better with more subtlety.

Style 2) Generic

You can just be yet another charity organization for yet another undifferentiated cause and try to get somewhere anyway. Some organizations have success with this. They aren't super important, they aren't super prestigious, but they put in the work and get somewhere.

SENS does some stuff in this direction (e.g. runs yet another small stakes matching fundraising), but isn't by any means great at it. For example the website isn't very well done, nor the subreddit, blog or newsletter.

Note, btw, that matching donation drives are bad and should not be done. See: **http://blog.givewell.org/2011/12/15/why-you-shouldnt-let-donation-matching-affect-your-giving/**

I tried explaining the problems with matching donations to "Reason" (the Fight Aging guy) at more length at the GRG email group but he was unwilling to address/discuss the problem.

Style 3) Reason

The third style is to focus on ideas and the intellect. Really seriously, not in the token way that's common. Here is one way to do this to give you the flavor:

Have high quality public discussions and challenge the entire public to offer any criticism of SENS, and answer every single criticism so you can honestly say there are literally no unanswered criticisms of SENS.

Saying that properly requires not just answering all the criticisms you know of, but also making a serious effort to seek them out in the first place, which

involves, for example, having discussion forum of some kind for people to post criticisms at where they expect to be heard and taken seriously. For criticism to be fully possible, you also have to answer questions so people can get you to take stances on every issue and potentially criticize your answers to the questions. They have to be able to draw out more claims from you and get things clarified.

This approach isn't just about telling people SENS is super important and intellectually correct, and acting the part. It also means SENS will get all kinds of ideas, suggestions, comments, feedback and criticism from the public. And some of it will be correct and SENS will learn something too. And it also means one member of the public can answer the question of another member of the public – there can be an interested group of people being helpful.

Broadly, I would say if people are too damn stupid and irrational and have no interest in thinking, SENS is pretty screwed anyway. But I don't think they all are, and I think you ought to try and give people the benefit of the doubt and stop treating them like they are beneath you. I think SENS ought to take the position that people really do have minds, and they matter – if they don't there honestly isn't much point in saving their lives anyway. Don't just ask for monetary donations, show you care about ideas by seeking them out too.

Note these 3 styles are incompatible. The prestige approach appeals to the irrational side of people. Focusing on reason isn't generic, it would stand out. Being generic isn't prestigious. So it's important to pick something and focus, rather than do a little of everything badly.

I recommend the Reason style because it's the only one where SENS is at an advantage. SENS does not have the most expertise at impressing fools with prestige, or at grassroots hard work and community building and running charities. And SENS has no inherent advantage at those activities. That SENS could save millions of lives, and has some good arguments for its importance, is only a major advantage intellectually. In the prestige and generic games, people with much worse causes will say they are important too or whatever else, and since there isn't an intellectual atmosphere they can get away with those claims.

I think SENS should focus on where it has a large advantage over almost all rivals. (I am not personally convinced SENS is the most important cause in the

world. But I agree it's a top cause, much better than the vast majority of causes.)

As a separate topic, consider that SENS would like a LOT of money. Like $100,000,000/yr for a decade. SENS, therefore, could use knowledge about money and economics. This kind of knowledge is necessary to use the money well. Consider that you wouldn't want an economically illiterate person deciding how to spend a million dollars. Well, at the billion dollar level, you wouldn't want a person with, say, "above average" economics knowledge either, you'd want world class knowledge to be involved. And it really helps to know how to deal with this money before asking for it, instead of telling people to trust that you'll figure it out correctly after getting it. And understanding these things is important for speaking intelligently to potential donors about these subjects.

This means, for example, familiarity with economics books such as *Capitalism: A Treatise on Economics* and *Human Action* (the best two major economics books). Preferably much more.

This does NOT mean that Aubrey should read those books. Understanding economics (not just reading a few books but studying it enough to really understand the material) is HARD and TIME CONSUMING. Therefore, it is an appropriate area for specialization and division of labor. SENS should have access to SOMEONE who knows this stuff, and who can relay important points to Aubrey and others when they are relevant.

Economics is not something everyone should learn, but it is important to basically everyone, and certainly to SENS which wants to deal with huge quantities of wealth. This is just like science: not everyone should be a scientist, division of labor is good, but science is important to everyone (and many organizations ought to have science advisors of some sort).

Similar lines of reasoning apply to quite a few other areas besides economics, such as epistemology (an understanding of the best methods of reasoning, and of philosophy of science, are two things that could aid SENS), moral philosophy (some of the objections to SENS involve moral issues), political philosophy (some actual and potential SENS projects involve the government), and computer science (maybe instead of preserving our bodies, we should upload our minds into computers. if we could accomplish that faster and cheaper than SENS, it could be the better option).

For each area, there are ongoing debates about which ideas in the field are right, which specialist experts are actually fools in disguise, which books are good, and so on. How is SENS to deal with this?

There is no way other than open rational public discussion. It leads back into the issue of discussion. Get a SENS economics expert who will address all public criticism, address all questions and issues about his economics claims, and so on. Open-ended rational discussion addressing all the issues is the only way to sort out the messes in all the various fields full of disagreement. I know this is hard and not SENS' expertise, but there is no way around it. This is what reason, truth-seeking and getting stuff right requires. The truth isn't easy to come by, too bad, suck it up and deal with it; there are no shortcuts.

SENS should not BET ITS FUTURE on the proposition that economics is irrelevant and ignorance of it won't lead to any major mistakes. Nor should SENS bet its future on siding with any particular side in the economics debates and not have that stance fully open to criticism and revision in case it's mistaken. And the same goes for other fields besides economics too.

SENS is struggling. It's badly underfunded. This stuff is URGENT and LIFE OR DEATH. SAY SO. CLEARLY. EVERYWHERE. Don't tell people everything is fine, tell the truth, it's NOT. Most current SENS communications act like these ideas about SENS' urgency are FALSE and actually everything is fine and not too urgent.

I think the most important thing is consistency. Have a consistent message and act commensurate with it. Have a consistent plan instead of a little from several styles.

I have more to say (lots), and more details for these points, but I think this is enough to get started. Please do not say "good points, you're very smart" and then proceed to do your (inevitable) initial misunderstandings of what I meant, without further discussion, in private (as is typical with this kind of thing).

PS Why didn't I write this sooner? Partly because of the contact form, as addressed above, and also the lack of any good SENS discussion place. Another major reason is b/c even now I don't really expect much to change, I don't expect this to have much effect. One reason is because I don't expect you guys to agree

with everything I say INITIALLY (which is completely fine and reasonable). And I don't expect you to discuss all this to resolution (which is problematic, it blocks **Paths Forward**, which is irrational). One reason for these low expectations is SENS does little to differentiate itself from all the other non-profits out there, and I certainly wouldn't expect most orgs to really listen to comments like these and make big changes.

But Aubrey asked me to write (some of) this, and anyway I think it's interesting. And SENS is important – as far as medical science, it impresses me more than anything else I've seen – so I hope this helps.

Update: I received a bad reply from Michael Rae and wrote some **comments** on it.

Update 2: I wrote **SENS Against Specialization and Division of Labor**.

SENS Against Specialization and Division of Labor

SENS has a budget of around 4 million dollars a year.

from this, they are unwilling to spend much or any on their website. (not sure the exact amount, i know they've asked for volunteers, and whatever they bought or didn't buy is low quality.)

i would strongly suspect they ARE willing to spend some money on an accountant, a lawyer, and perhaps a few other non-SENS-specific functions. as well they should be.

they also should spend money on a website. it's not very hard to buy quality web knowledge and work. it's readily available on the market at prices very low compared to the value provided, and easily affordable on their budget.

this is something many other organizations do. it's not a weird FI-only idea. SENS is frankly just plain incompetent here.

there are some other areas where SENS is making similar errors which are less well understood in general, and where useful expertise is less readily available to purchase.

if you want a good website, you can have that set up tomorrow. it's no problem at all to find a person or group. if you want a GREAT website, you should shop around some, but it's not that hard.

what if you want economics expertise? SENS deals with quite a bit of money – around 4 million a year. that's enough that i think they should spend more than $0/yr on economics expertise (at least if they could find some to hire – which i strongly suspect is completely possible despite the market for it being more problematic than for websites).

further, SENS wants to deal with at least 100 million a year. they have openly and explicitly asked the public for that amount as a minimum for the project they

regard as most important (robust mouse rejuvenation). and they want that 100 million budget for 10 years or more. that is a LOT of money. if 4 million a year is too trivial to merit more than $0 of economics knowledge (i disagree!!!), surely 100 million a year has room in the budget for economics expertise. yet i don't believe SENS would hire economics expertise even at that budget level. they expressed serious hostility to this kind of thinking. they don't see why people dealing with huge quantities of money would need to know anything about money. additionally, i pointed out that they ought to understand how to use the budget they request BEFORE requesting it, which they were also hostile to.

but actually SENS already has some economics knowledge. everyone who works at SENS knows SOMETHING about economics. it is amateur level knowledge. they are dabblers. they think that's good enough. they think they are clever enough to get by, and/or economics is easy, and/or what's well known about economics is all they need to know and knowing anything more would be pointless. that is very foolish.

suppose, hypothetically, that Aubrey de Grey (AdG) is smarter than anyone working in the field of economics. and suppose that AdG puts an equivalent of 2 hours a month of his SENS work into thinking related to economics issues. this is completely plausible. he thinks about money, how to get money, different places money comes from, what to do with money, and so on.

what are the consequences?

nothing but disaster, even though, by premise, AdG is smarter than any economist.

first, AdG is by far the best person to do some tasks – such as explain SENS on podcasts. the consequences are either to do without that, or to have someone worse at it do it. either it's going to be done 2 hours less per month, or someone lesser to the amazing genius AdG would be doing it in his place – a huge loss. the only way this SENS podcast advocacy would not be lost is if there is something even more important AdG is giving up instead – something where to an even greater extent than SENS podcasting, AdG is the best suited to do it – in which case if he freed up 2 hours per month it would go to that even more important task instead.

second, AdG is not an economics specialist. being the smartest person in the world could not make up for this. why? because the more time you spend on economics, the more you can specialize in the field. if you only work on economics 2 hours a month, for SENS, that will justify very little or no time spent reading economics books. but a specialist, who does economics work for 100 hours per month, could very reasonably also devote 20 hours per month to reading economics books. this is a huge advantage which more than makes up for AdG being the smarter clever person in general. additionally, during those 100 hours per month of economics work, the specialist will gain benefits too. he'll get accustomed to many common economics problems and get practice at solving them quickly. all that practice and experience and familiarity will help. and the specialist will keep up-to-date better than the non-specialist, because he does frequent work in the field which will benefit from staying up-to-date. and the specialist will be able to have discussions where he challenges his views about economics, tests them in debate, listens to people with new ideas, and so on. why will he find time for those things? because he spends so 100 hours per month doing economics work, any little improvement in his craft will be 50 times as valuable to him as it will be to AdG who spends 2 hours per month. (and actually the difference is larger, because a specialist is expected to know his field, and will care about his reputation in the field, whereas AdG will be recognized as wearing many hats, and barely dealing with economics, and will therefore be forgiven for not doing it as well as a specialist would be expected to.)

so there is a double issue. AdG would be giving up time to do what he's better at than economics – doing the stuff where is able to get the most valuable work done per hour – and he would also be at a huge disadvantage due to not specializing in economics.

and even if AdG was so great he could do economics work equally well, and twice as fast, as an economist, he STILL shouldn't do it. because his advantage at SENS work is even larger than that. if AdG can do SENS-specific work three times as well as the next best person, and economics work twice as well, then he should only do SENS work and hire an economist (for twice the number of hours it'd take AdG). That beats having to hire someone to do SENS work in place of AdG for three times the number of hours!

put another way: suppose AdG can create $300 per hour of value doing SENS work, or $200 per hour of value doing economics. i think the real ratio is more

like 100 to 1, rather than 1.5 to 1, but this will illustrate my point. And suppose if AdG hires people to do these things instead of him, the best people he can find aren't as good as him – they can create $100 of value per hour for SENS work or economics work. Then very simply, AdG should not do economics work – he's better off outsourcing that, even though he's (hypothetical) the best in the world at it, because his advantage at SENS work is even greater. he is relatively more productive when doing SENS work over economics work. and other people are equally productive. (more realistically, SENS is obscure and economics is common, so other people in general would be relatively more productive at economics work over SENS work, which would only increate the advantage of AdG sticking to SENS work).

this last point i've explained is a well known economics concept called "comparative advantage".

if you ask AdG if he knows what comparative advantage is, and how it works, my guess is that he does. yet i still think it's important to hire an economics specialist to help advise on topics including comparative advantage. why? because there are different senses of understanding comparative advantage.

a specialist would have an ACTIVE understanding of comparative advantage – he will have used the concept many times in many different situations. he will be able to recognize, pro-actively, many times he'd be able to use it. he'll have experience stretching it to use in all kinds of cases where it doesn't obviously apply.

someone like AdG, who spends little time on economics, would have a PASSIVE understanding of comparative advantage. he would be able to tell you what it is IF YOU ASK HIM. he might bring it up himself in a few situations – especially if you asked him about international trade between countries, especially countries where one is at a big advantage (e.g. industrial first world country trading with a third world poor country). That's the best known context for thinking about comparative advantage, and the most common one discussed when the concept taught. But AdG hasn't read books about all the other situations comparative advantage is relevant to, he hasn't practiced finding ways to use it in many situations. His way of knowing what it is if you ask is completely different than superior sort of understanding that a specialist would have.

so even when AdG thinks, “oh i’ve got this, i know what comparative advantage is, there’s no need for an economics specialist to tell me that” he would be wrong.

there is no way SENS gets by with an actual expense of $0 on economics. it is relevant to what they do. they must think about it some. depending on their ideas about economics, it would to some extent lead them to different strategies. and AdG discusses economics in his book *Ending Aging* very literally – he tries to explain his ideas about the effect on the country, economy (including medical prices), government, and world if everyone had AIDs and we had to produce enough AIDs medicine for everyone. That is very clearly partly an economics issue.

so AdG and/or others at SENS, who are not economics specialists, inefficiently do some economics work, instead of sticking to SENS-specific work that they are, relatively, better at doing. and i think they make some large mistakes due to their arrogance to do work outside their fields. and they are completely hostile to the idea that maybe they should spend more than $0 getting specialist help with economics, rather than sacrificing SENS-specific work to dabble in it themselves. the people at SENS may be pretty smart, but there are very smart people working on economics too, and it’s HARD even for people who study it extensively and specialize in it. it’s completely unrealistic and unreasonable for SENS to be like, “ok we’re doing the most important thing in the world. now for this AIDS hypothetical, and some other matters, let’s try amateur hour. we can probably get away with that. it’ll be fine. and it doesn’t require any humility or respect for other people who aren’t doing what is obviously the most important work in the world.”

all of what i’ve said applies to other topics besides economics. they dabble in many other areas: philosophy of critical thinking, philosophy of science, philosophy of persuasion, political philosophy (they have various ideas about the government and its agencies, and how to deal with them and talk about them), and some rather different fields like how to run a charity fundraiser (an area where they have made big mistakes such as using **matching donation fundraising**). and what about marketing? they appear completely clueless about that. it’s ridiculous that they don’t have a specialist guiding them to do a much better job with marketing. i’ll let Steve Jobs explain this one:

Becoming Steve Jobs: the evolution of a reckless upstart into a visionary leader by Brent Schlender and Rick Tetzeli:

> [Context: Seva is a philanthropy type foundation. They are having a meeting, at the start, about how the make the world better. One of the guys had just been a significantly involved in eradicating smallpox in India. Now they want to make Seva and do more. What would be the best thing to work on? They decide on curing blind people in the third world.]
>
> [Steve Jobs] sat down and started listening. The decision to create a foundation had already been made; the question now on the table was how to tell the world about Seva, its plans, and the men and women who would implement those plans. Steve found most of the ideas embarrassingly naïve. The discussion seemed more appropriate for a PTA meeting; at one point, everyone but Steve heatedly debated the finer points of a pamphlet they wanted to create. A pamphlet? That's the best these people could dream up? These so-called experts may have achieved notable progress in their own countries, but here they were clearly out of their league. Having a grand, bold goal was useless if you didn't have the ability to tell a compelling story about how you'd get there. That seemed obvious.
>
> As the discussion meandered, Steve found his own attention wandering. "He had walked into that room with his persona from the Apple board meeting," Brilliant remembers, "but the rules for doing things like conquering blindness or eradicating smallpox are quite different." From time to time he'd pipe up, but mostly to interject a snide remark about why this or that idea could never fly. "He was becoming a nuisance," says Brilliant. Finally, Steve couldn't take it anymore. He stood up.
>
> "Listen," he said, "I'm telling you this as someone who knows a thing or two about marketing. We've sold nearly a hundred thousand machines at Apple Computer, and when we started no one knew a thing about us. Seva is in the same position Apple was in a couple of years ago. The difference is you guys don't know diddly about marketing. So if you want to really do something here, if you really want to make a difference in the world and not just putter along like every other nonprofit that people have never heard of, you need to hire this guy named Regis McKenna—he's the king of marketing. I can get him in here if you'd like. You should have the best. Don't settle for second best."

The result? They made Steve Jobs cry (yes, literally) and kicked him out of the meeting (yes, literally). (And then, I take it, did a much worse job fighting blindness than they could have). That's how hostile and unreasonable they were. They wanted to do this extremely important humanitarian work (their own view), but they absolutely would not consider hiring some world class expertise to do it right.

And SENS, which claims to be basically the most important thing in the world, and which has enough money to hire help, won't hire top experts either – be it about economics, marketing, philosophy, fundraising, or even making a good website.

By the way, I'm not even going to send AdG a link to this, even though we had a long discussion before. I wrote to him to tell him I'd given up on SENS – and why. He did not reply. He is too unreasonable to talk to, or tell things like this. He won't listen. I think it's hopeless. It's a ridiculous situation. I may well literally die because AdG won't listen, and yet he convinced me to give up (I just had a some thoughts I wanted to write down, because it's interesting and I think about things like this, but in another month maybe I'll forget about SENS).

I could fucking cry.

Steve Jobs apologized to Seva for trying to help. At least I won't be apologizing to SENS.