# A Whirlwind Tour Of LW Rationality

## 6 Books In 32 Pages

Version 1.0.0

Written by JBeshir (Viewable online at https://summaries.beshir.org/lw-tour)

# A Whirlwind Tour Of LW Rationality: 6 Books In 32 Pages

This is an attempt to summarise the propositions of the online 'rationalist' community, originally centred around Overcoming Bias and LessWrong, now largely dispersed to various communities and organisations like Slatestarcodex, the Center for Applied Rationality, the Machine Intelligence Research Institute, and the Effective Altruism movement, amongst others.

I have held off on writing this in the past out of a suspicion that I would not do it justice, but have decided that it is better done badly than not at all. My apologies to Eliezer Yudkowsky for mangling their work. I have reservations in parts, but I agree with or find it plausible in gist.

The structure is that of a whirlwind tour, with little narrative beyond ordering of the propositions, with citations to the source post for each, to permit drilling down into interesting or contentious parts and reading existing critique by the community.

This is to enable useful examination of the ideas and their assumptions by people who have things to do other than reading millions of words on the topic, to permit those who have picked up ideas from the community to see their surrounding context and related ideas, and to serve as an index to enable those who disagree to identify their points of departure.

## Hurrying Along, What Is "LW Rationality"?

- An empiricist, methodologically reductionist, materialist, atheist set of beliefs about epistemology, decision theory, cognition in general, and human cognition in particular, with proposed limitations and common errors resulting from the use of imperfect heuristics.
- A set of beliefs about how reductionism and materialism are grounded in epistemology.
- A set of beliefs about human values, in particular the belief that our true preferences are consequentialist, and that we pursuit our preferences ineffectively.
- A partial set of strategies for mitigating or avoiding proposed errors in human cognition.
- A very partial set of strategies for more effectively achieving our values.
- A coined jargon of labels for these beliefs, limitations, errors, and strategies, used to reference them quickly and debate them and their further implications.

So, roughly a mixture of analytic philosophy and pop cognitive science. The basic attitude to human cognition is that of Kanheman's Thinking Fast And Slow, which I recommend. The consensus reference to LW rationality itself is Eliezer Yudkowsky's core Sequences, blog posts with examples and stories of a transhumanist, speculative flavour. They have since been collected into the book Rationality: AI to Zombies, which is available for free and is the best place to start if seeking a fuller understanding of the propositions here. A description of how it compares to and connects with academia, with references to related works and research, was written by lukeprog.

Some parts of these are relatively well accepted; others proved controversial with the community. The tour follows, roughly a page per sequence, using the book's ordering of the sequences.

## Map and Territory A: Predictably Wrong

*Epistemic rationality* is using new evidence to improve the correspondence between your mental map and the world. *Instrumental rationality* is effectively accomplishing your goals.[1]

Rationality does not conflict with having strong feelings about true aspects of the world.[2]

Epistemic rationality is useful if you are curious, if you want to be effective, or if you regard it as a moral duty, the last of which can be problematic.[3] A *bias* is an obstacle to epistemic rationality produced by the 'shape' of our mental machinery. We should be concerned about any obstacle.[4]

We use an *availability heuristic* to judge the probability of something by how easily examples of it come to mind. This is imperfect, creating the *availability bias*. Selective reporting is a major cause.[5]

We use a *judgement of representativeness* to judge the probability of something by how typical it sounds. This suffers from the *conjunctive bias*, where adding more details increases perceived probability.[6]

We tend to examine only the scenario where things go according to plan. This suffers from the *planning fallacy*, in which difficulties and delays are underestimated.[7]

We use our own understanding of words to evaluate how others will understand them. This underestimates differences in interpretation, leading to the *illusion of transparency*.[8]

*Inferential distance* is the amount of explanation needed to communicate one person's reasoning to another. We routinely underestimate it. This is because our background knowledge differs a lot more now than it used to in the past.[9]

A metaphor for the human brain is a flawed lens that can see its own flaws.[10]

---

[1] http://lesswrong.com/lw/31/what_do_we_mean_by_rationality/
[2] http://lesswrong.com/lw/hp/feeling_rational/
[3] http://lesswrong.com/lw/go/why_truth_and/
[4] http://lesswrong.com/lw/gp/whats_a_bias_again/
[5] http://lesswrong.com/lw/j5/availability/
[6] http://lesswrong.com/lw/jk/burdensome_details/
[7] http://lesswrong.com/lw/jg/planning_fallacy/
[8] http://lesswrong.com/lw/ke/illusion_of_transparency_why_no_one_understands/
[9] http://lesswrong.com/lw/kg/expecting_short_inferential_distances/
[10] http://lesswrong.com/lw/jm/the_lens_that_sees_its_flaws/

## Map and Territory B: Fake Beliefs

A belief should be something that tells you what you expect to see; it should be an *anticipation-controller*.[11]

Taking on a belief can acquire social implications, and this results in a variety of compromises to truth seeking.[12] It is possible to believe you have a belief while still truly expecting to see the opposite; this is *belief-in-belief*.[13][14] Holding a neutral position on a question is a position on it like any other.[15]

Religious claims to be non-disprovable metaphor are a socially-motivated backdown from what were originally beliefs about the world, with claims to ethical authority remaining because they have not become socially disadvantageous.[16] At other times, we can see socially-motivated claims of extreme beliefs, as a way to cheer for something.[17]

*Belief as attire* is belief that is professed in order to show membership of a group.[18]

Some statements exist simply to tell the audience to applaud and do not actually express any belief; we call these *applause lights*.[19]

## Map and Territory C: Noticing Confusion

When uncertain, we want to focus our anticipation into the outcome which will actually happen as much as possible.[20]

It means exactly what you think it means for a statement to be true. *Evidence* is an event, entangled by cause and effect, with what you want to know about. Things that react to that event can become entangled with what you want to know about in turn. Beliefs should be determined in a way that makes them entangled, as this is what makes them accurate. You must be conceivably able to believe otherwise given different observations.[21]

Scientific evidence and legal evidence are subsets of rational evidence.[22]

---

[11] http://lesswrong.com/lw/i3/making_beliefs_pay_rent_in_anticipated_experiences/
[12] http://lesswrong.com/lw/gt/a_fable_of_science_and_politics/
[13] http://lesswrong.com/lw/i4/belief_in_belief/
[14] http://lesswrong.com/lw/i5/bayesian_judo/
[15] http://lesswrong.com/lw/yp/pretending_to_be_wise/
[16] http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/
[17] http://lesswrong.com/lw/i6/professing_and_cheering/
[18] http://lesswrong.com/lw/i7/belief_as_attire/
[19] http://lesswrong.com/lw/jb/applause_lights/
[20] http://lesswrong.com/lw/ia/focus_your_uncertainty/
[21] http://lesswrong.com/lw/jl/what_is_evidence/
[22] http://lesswrong.com/lw/in/scientific_evidence_legal_evidence_rational/

The amount of entanglement needed to justify a strong belief depends on how improbable the hypothesis was to begin with, which is related to the number of possible hypotheses.[23][24]

*Occam's Razor* is the principle that the correct explanation is the simplest that fits the facts. The simplest explanation must be defined as the shortest length it takes to fully specify a program that simulates the explanation/a universe that performs the explanation rather than English sentence length. *Solomonoff Induction* is a formalisation of this; one variant predicts sequences by assigning a base probability to programs of $2^{-<\text{bit length}>}$ and then weights based on how their predictions fit. This definition reduces probability of an explanation equally to the extent to which it simply embeds a copy of the observations, and only so rewards explanations which are compressed relative to the observations.[25]

Your strength as a rationalist is your ability to notice confusion; your sense that your explanation feels forced.[26]

Absence of evidence is evidence of absence. If something being present increases your probability of a claim being true, then its absence must decrease it, in amounts depending on how likely the presence was in either case.[27] There is *conservation of expected evidence*.[28]

We have a *hindsight bias* which makes us think we already believed something when we read it.[29]

## Map and Territory D: Mysterious Answers

A *fake explanation* is an explanation that can explain any observation.[30] Using scientific-sounding words in one is using *science as attire*, and not actually adhering to science.[31] After seeing a thing happen, we tend to come up with explanations for how it was caused by a phenomenon, even when we couldn't have predicted it ahead of time from our knowledge of that phenomenon. This is *fake causality* and made hard to notice by the hindsight bias. The hindsight bias is caused by failing to exclude the evidence we get from seeing a claim when evaluating how likely we thought it was before we saw it.[32]

---

[23] http://lesswrong.com/lw/jn/how_much_evidence_does_it_take/
[24] http://lesswrong.com/lw/jo/einsteins_arrogance/
[25] http://lesswrong.com/lw/jp/occams_razor/
[26] http://lesswrong.com/lw/if/your_strength_as_a_rationalist/
[27] http://lesswrong.com/lw/ih/absence_of_evidence_is_evidence_of_absence/
[28] http://lesswrong.com/lw/ii/conservation_of_expected_evidence/
[29] http://lesswrong.com/lw/im/hindsight_devalues_science/
[30] http://lesswrong.com/lw/ip/fake_explanations/
[31] http://lesswrong.com/lw/ir/science_as_attire/
[32] http://lesswrong.com/lw/is/fake_causality/

*Positive bias* is attempting to confirm rather than disconfirm theories, which fails to properly test them.[33]

There is a normal human behaviour when asked to proffer an explanation where we pull out phrases and offer them without a coherent model. We call this *guessing the teacher's password*.[34] A good way to examine whether you truly understand a fact rather than have it memorised as a password answer is to ask whether you could regenerate it if forgotten.[35]

It is not necessary to counter irrationality with irrationality, or randomness with randomness, despite this being the intuitive thing to do as a human.[36]

A fake explanation often serves as a sign to end further examination despite containing no real understanding, in which case it is a *semantic stopsign* or *curiosity-stopper*.[37] We should not expect answers to be 'mysterious', even for 'mysterious questions', such as the cause of fire or life.[38] Any time humans encounter a phenomenon, they can choose to try to explain it, worship it, or ignore it.[39]

The term 'emergence' is a contemporary fake explanation and semantic stopsign.[40] The word 'complexity' in the sense of a desired addition can also be so. It is tempting to assign fake explanations to mysterious parts when trying to understand something. This must be resisted.[41]

Eliezer failed at this in his earlier days, despite knowing to reject the standard 'fake explanations'; it takes a lot of improvement to not simply find new, interesting mistakes instead of the old ones.[42] Solving a mystery should make it feel less confusing, but it is difficult to learn what believing the old fake explanations felt like to recognise new ones.[43] Trying to visualise believing in ideas like "elan vital", without being able to immediately see your error, may help.[44]

Explanations like 'Science' can serve as curiosity-stoppers, by telling us that someone else knows the answer.[45]

---

[33] http://lesswrong.com/lw/iw/positive_bias_look_into_the_dark/
[34] http://lesswrong.com/lw/iq/guessing_the_teachers_password/
[35] http://lesswrong.com/lw/la/truly_part_of_you/
[36] http://lesswrong.com/lw/vo/lawful_uncertainty/
[37] http://lesswrong.com/lw/it/semantic_stopsigns/
[38] http://lesswrong.com/lw/iu/mysterious_answers_to_mysterious_questions/
[39] http://lesswrong.com/lw/j2/explainworshipignore/
[40] http://lesswrong.com/lw/iv/the_futility_of_emergence/
[41] http://lesswrong.com/lw/ix/say_not_complexity/
[42] http://lesswrong.com/lw/iy/my_wild_and_reckless_youth/
[43] http://lesswrong.com/lw/iz/failing_to_learn_from_history/
[44] http://lesswrong.com/lw/j0/making_history_available/
[45] http://lesswrong.com/lw/j3/science_as_curiositystopper/

# How To Actually Change Your Mind E: Overly Convenient Excuses

Humility is a complicated virtue, and we should judge it by whether applying it makes us stronger or weaker, and by whether it is an excuse to shrug. To be correctly humble is to take action in anticipation of one's own errors.[46]

A *package deal fallacy* is where you assume things traditionally grouped together must always be so. A *false dilemma* is presenting only two options where more exist. Justifications for noble lies are usually one of the two; it is preferable to seek a third alternative, which may be less convenient.[47]

Human hope is limited and valuable, and the likes of lotteries waste it.[48][49] There is a bias in which extremely tiny chances are treated as more than tiny in implication, and justify proclaiming belief in them. There is a tendency to arbitrarily choose to 'believe' or not believe a thing rather than reacting to probabilities.[50]

The *fallacy of grey* is to regard all imperfection and all uncertainty as equal. Wrong is relative.[51] There is a sizeable inferential distance from thinking of knowledge as absolutely true to understanding knowledge as probabilistic.[52] Eliezer says he would be convinced that 2 + 2 = 3 by the same processes that convinced him that 2 + 2 = 4; a combination of physical observation, mental visualization, and social agreement, such as observing that putting two more objects down beside two objects produced three objects.[53]

Because of how evidence works, a probability of 100% or 0% corresponds to infinite certainty, and requires infinite evidence to correctly attain. As a result it is always incorrect.[54] 0 and 1 are [in a sense] not probabilities.[55]

It is reasonable to care how other humans think, as part of caring about how the future and present look. This is somewhat dangerous, and so must be tempered by a solid commitment to respond to bad thinking only with argument.[56]

---

[46] http://lesswrong.com/lw/gq/the_proper_use_of_humility/
[47] http://lesswrong.com/lw/hu/the_third_alternative/
[48] http://lesswrong.com/lw/hl/lotteries_a_waste_of_hope/
[49] http://lesswrong.com/lw/hm/new_improved_lottery/
[50] http://lesswrong.com/lw/ml/but_theres_still_a_chance_right/
[51] http://lesswrong.com/lw/mm/the_fallacy_of_gray/
[52] http://lesswrong.com/lw/mn/absolute_authority/
[53] http://lesswrong.com/lw/jr/how_to_convince_me_that_2_2_3/
[54] http://lesswrong.com/lw/mo/infinite_certainty/
[55] http://lesswrong.com/lw/mp/0_and_1_are_not_probabilities/
[56] http://lesswrong.com/lw/hn/your_rationality_is_my_business/

# How To Actually Change Your Mind F: Politics and Rationality

*Politics is the mind-killer*. People cannot think clearly about politics close to them. In politics, *arguments are soldiers*. When giving examples, it is tempting to use contemporary politics. Avoid this if possible. If you are discussing something innately political, use an example from historic politics with minimal contemporary implications if possible.[57]

*Policy debates should not appear one-sided*. Actions with many consequences should not be expected to have exclusively positive or negative consequences. If they appear to, this is normally the result of bias. They may legitimately have lopsided costs and benefits.[58]

Humans tend to treat debates as a contest between two sides, where any weakness in one side is a gain to the other and visa versa, and whoever wins is correct on everything and whoever loses is wrong on everything. This is correct behaviour for a single, strictly binary question, but an error for any more complicated debate.[59]

The *fundamental attribution error* is a tendency in people to overly attribute the actions of others to innate traits, while overly attributing their own actions to circumstance as opposed to differences in themselves. Most people see themselves as normal.[60] Even your worst enemies are not innately evil, and usually view themselves as the heroes of their own story.[61]

Stupidity causes more random beliefs, not reliably wrong ones, so reversing the beliefs of the foolish does not create correct beliefs; *reversed stupidity is not intelligence*. Foolish people disagreeing does not mean that you are correct.[62]

Authority can be a useful guide to truth before you've heard arguments, but is not so after arguments.[63] The more distant from the specific question evidence is, the weaker it is. You should try to answer questions using direct evidence- *hug the query*. Otherwise learning abstract arguments, including about biases, can make you less rather than more accurate.[64]

Speakers may manipulate their phrasing to alter what aspects of a situation are noticed.[65] Simplifying language interferes with this, and allows you to recognise errors in your own speech.[66]

---

[57] http://lesswrong.com/lw/gw/politics_is_the_mindkiller/
[58] http://lesswrong.com/lw/gz/policy_debates_should_not_appear_onesided/
[59] http://lesswrong.com/lw/h1/the_scales_of_justice_the_notebook_of_rationality/
[60] http://lesswrong.com/lw/hz/correspondence_bias/
[61] http://lesswrong.com/lw/i0/are_your_enemies_innately_evil/
[62] http://lesswrong.com/lw/lw/reversed_stupidity_is_not_intelligence/
[63] http://lesswrong.com/lw/lx/argument_screens_off_authority/
[64] http://lesswrong.com/lw/ly/hug_the_query/
[65] http://lesswrong.com/lw/jc/rationality_and_the_english_language/
[66] http://lesswrong.com/lw/jd/human_evil_and_muddled_thinking/

# How To Actually Change Your Mind G: Against Rationalization

Because humans are irrational to start with, more knowledge can hurt you. Knowledge of biases gives you ammunition to use against arguments, including knowledge of this one.[67]

Expect occasional opposing evidence for any imperfectly exact model. You should not look for reasons to reject it, but *update incrementally* as it suggests. If your model is good, you will see evidence supporting it soon.[68] You should not decide what direction to change your opinion in by comparing new evidence to old arguments; this double-counts evidence.[69]

The sophistication with which you construct arguments does not improve your conclusions; that requires choosing what to argue in a manner that entangles your choice with the truth.[70]

Reaction to evidence that someone is filtering must include reacting to knowledge of the filtering. Knowing what is true can require looking at evidence from multiple parties.[71]

*Rationalization* is determining your reasoning after your conclusion, and runs in the opposite direction to rationality.[72] You cannot create a rational argument this way, whatever you cite.[73]

Humans tend to consider only the critiques of their position that they know they can defeat.[74] A *motivated skeptic* asks if the evidence compels them to believe; a *motivated credulist* asks if the evidence allows them to believe. *Motivated stopping* is ceasing the search for opposing evidence earlier when you agree, and *motivated continuation* is searching longer when you don't.[75]

*Fake justification* is searching for a justification for a belief which is not the one which led you to originally hold it.[76] Justifications for rejecting a proposition are often not the person's *true objection*, which when dispelled would result in the proposition being accepted.[77]

---

[67] http://lesswrong.com/lw/he/knowing_about_biases_can_hurt_people/
[68] http://lesswrong.com/lw/ij/update_yourself_incrementally/
[69] http://lesswrong.com/lw/ik/one_argument_against_an_army/
[70] http://lesswrong.com/lw/js/the_bottom_line/
[71] http://lesswrong.com/lw/jt/what_evidence_filtered_evidence/
[72] http://lesswrong.com/lw/ju/rationalization/
[73] http://lesswrong.com/lw/jw/a_rational_argument/
[74] http://lesswrong.com/lw/jy/avoiding_your_beliefs_real_weak_points/
[75] http://lesswrong.com/lw/km/motivated_stopping_and_motivated_continuation/
[76] http://lesswrong.com/lw/kq/fake_justification/
[77] http://lesswrong.com/lw/wj/is_that_your_true_rejection/

Facts about reality are often entangled with each other.[78][79] Maintaining a false belief often requires other false beliefs, including deception about evidence and rationality themselves.[80]

## How To Actually Change Your Mind H: Against Doublethink

In *doublethink*, you forget then forget you have forgotten. In *singlethink*, you notice yourself forgetting an uncomfortable thought and recall it.[81]

If you watch the risks of doublethink enough to do it only when useful, you cannot do it. If you do not, you will do it where it harms you. Doublethink is either not an option or harmful.[82]

The above on doublethink not be a dispassionate reporting of the facts; Eliezer admits that they may have been tempted into trying to create a self-fulfilling prophecy. They then say that it may be wise to at least tell yourself that you can't self-deceive, so that you aren't tempted to try.[83]

It is possible to lead yourself to think you believe something without believing it. Believing that a belief is good can lead you to false belief-in-belief.[84][85] We often do not separate believing a belief from endorsing a belief. Belief-in-belief can create apparently contradictory beliefs.[86]

## How To Actually Change Your Mind I: Seeing With Fresh Eyes

*Anchoring* is a behaviour in which we take a figure we've recently seen and adjust it to answer questions, making results depend on the initial anchor. A strategy for countering it might be to dwell on an alternative anchor if you notice an initial guess is implausible.[87]

*Priming* is an aspect of our brain's architecture. Concepts related to ideas we've recently had in mind are recalled faster. This means that completely irrelevant observations influence estimates and decisions. This is known as *contamination*. It supports *confirmation bias*; having an idea in our head makes compatible ideas come to mind more easily, making us more receptive to confirming than disconfirming evidence for our beliefs.[88]

Some evidence suggests that we tend to initially believe statements, then adjust to reject false ones. Being distracted makes us more likely to believe statements explicitly labeled as false.[89]

---

[78] http://lesswrong.com/lw/uw/entangled_truths_contagious_lies/
[79] http://lesswrong.com/lw/9a/of_lies_and_black_swan_blowups/
[80] http://lesswrong.com/lw/uy/dark_side_epistemology/
[81] http://lesswrong.com/lw/k0/singlethink/
[82] http://lesswrong.com/lw/je/doublethink_choosing_to_be_biased/
[83] http://lesswrong.com/lw/1o/dont_believe_youll_selfdeceive/
[84] http://lesswrong.com/lw/r/no_really_ive_deceived_myself/
[85] http://lesswrong.com/lw/s/belief_in_selfdeception/
[86] http://lesswrong.com/lw/1f/moores_paradox/
[87] http://lesswrong.com/lw/j7/anchoring_and_adjustment/
[88] http://lesswrong.com/lw/k3/priming_and_contamination/
[89] http://lesswrong.com/lw/k4/do_we_believe_everything_were_told/

The *hundred-step rule* is the principle that because neurons in the human brain are slow, any hypothesised operation can be very parallel but must complete in under a hundred sequential neuron spikes. It is a good guess that human cognition consists mostly of cache lookups.

We incorporate the thoughts of others into this cache, and alone could not regenerate all the ideas we've collected in a single lifetime. We tend to incorporate and then repeat or act on *cached thoughts* without thinking about their source or credibility.[90]

"Outside the box" thinking is a box of its own, and along with stated efforts at originality and subversive thinking follows predictable patterns; genuine originality requires thinking.[91] When a topic seems to have nothing to be said, it can mean we do not have any related cached thoughts, and find generating new ones difficult.[92]

The events of history would sound extremely strange described to someone prior to them.[93] We tend to treat fiction as history which happened elsewhere. This causes us to favour hypotheses which fit into fun narratives, over other hypotheses that might be likely.[94]

A model which connects all things contains the same information as a model that connects none. Information is contained in selectiveness about connections, and the more fine-grained this is the more information is contained. The *virtue of narrowness* is the definition and use of narrow terms and ideas rather than broad ones.[95]

One may sound deep by coherently expressing cached thoughts that the listener hasn't heard yet. One may be deep by attempting to see for yourself rather than following standard patterns.[96]

We change our mind less often than we think, and are resistant to it. A technique to mitigate against this is to *hold off on proposing solutions* as long as possible.[97][98]

Because of confirmation bias, we should be suspicious of ideas that originally came from sources whose output was not entangled with the truth. However, to disregard other evidence entirely in favour of judging the original source would be the *genetic fallacy*.[99]

---

[90] http://lesswrong.com/lw/k5/cached_thoughts/
[91] http://lesswrong.com/lw/k6/the_outside_the_box_box/
[92] http://lesswrong.com/lw/k7/original_seeing/
[93] http://lesswrong.com/lw/j1/stranger_than_history/
[94] http://lesswrong.com/lw/k9/the_logical_fallacy_of_generalization_from/
[95] http://lesswrong.com/lw/ic/the_virtue_of_narrowness/
[96] http://lesswrong.com/lw/k8/how_to_seem_and_be_deep/
[97] http://lesswrong.com/lw/jx/we_change_our_minds_less_often_than_we_think/
[98] http://lesswrong.com/lw/ka/hold_off_on_proposing_solutions/
[99] http://lesswrong.com/lw/s3/the_genetic_fallacy/

# How To Actually Change Your Mind J: Death Spirals and the Cult Attractor

The *affect heuristic* is when subjective impressions of goodness/badness act as a heuristic. It causes the manner in which a problem is stated and irrelevant aspects of a situation to change the decisions we make.[100] The *halo effect* is this applied to people; when our subjective impression of a person in one regard, such as appearance, alters our judgement of them in others.[101]

We overestimate the altruism of those who run less risk compared to those who run more, and attribute less virtue to people who are generous for lesser as well as greater need.[102] We lionize messiahs for whom doing great things is easy over those for whom it is hard.[103]

We tend to evaluate things against nearby points of comparison.[104] When we lack a bounded scale to put our estimates within, we make one up, inconsistently between people.[105]

An *affective death spiral* is a scenario in which a strong positive impression assigned to one idea causes us to improve our impressions of related ideas, which we then treat as confirmation of the original idea in a self-sustaining cycle.[106] We can diminish the effect of positive impressions enough to prevent this by splitting big ideas into smaller ones we treat independently, reminding ourselves of the conjunctive bias and considering each additional claim to be a burdensome detail, and following the suggestions in the Against Rationalization sequence.[107]

Considering it morally wrong to criticise an idea accelerates an affective death spiral.[108] *Evaporative cooling of group beliefs* is a scenario in which as a group becomes more extreme, moderates leave, and as they are no longer acting as a brake, the group becomes yet more extreme, in a cycle. This is another reason why tolerating dissent is important.[109]

A *spiral of hate* is the mirror image of an affective death spiral, in which a strong negative impression of a thing causes us to believe related negative ideas, which we then treat as

---

[100] http://lesswrong.com/lw/lg/the_affect_heuristic/
[101] http://lesswrong.com/lw/lj/the_halo_effect/
[102] http://lesswrong.com/lw/lk/superhero_bias/
[103] http://lesswrong.com/lw/ll/mere_messiahs/
[104] http://lesswrong.com/lw/lh/evaluability_and_cheap_holiday_shopping/
[105] http://lesswrong.com/lw/li/unbounded_scales_huge_jury_awards_futurism/
[106] http://lesswrong.com/lw/lm/affective_death_spirals/
[107] http://lesswrong.com/lw/ln/resist_the_happy_death_spiral/
[108] http://lesswrong.com/lw/lo/uncritical_supercriticality/
[109] http://lesswrong.com/lw/lr/evaporative_cooling_of_group_beliefs/

strengthening the original impression. You can correspondingly observe it become morally wrong to urge restraint or to object to a criticism. It, too, leads to poor choice of action.[110]

Humans, once divided into opposing groups, will naturally form positive and negative stereotypes of the two groups and engage in conflict.[111] Every cause has a natural tendency for its supporters to become focused on defending their group, even if they declare 'rationality' to be their goal.[112]

Beware being primarily a guardian of the truth rather than primarily a seeker of it.[113] The Nazis can be understood as would-be guardians of the gene pool.[114]

There are things we know now which earlier generations could not have known, which means that from our perspective we should expect elementary errors even in our historic geniuses. This is a defining attribute of scientific disciplines. It feels unfair to consider things they could not have known to be flaws in their ideas, but nevertheless they are. It is foolish to declare a system of ideas to be closed to further development. We already have examples of people who declared themselves to be about being Rational who fell into that trap in history.[115]

Two ideas for countering a tendency towards affective death spirals around a group are to prefer using and describing techniques over citing authority, and to deliberately look foolish to reduce the positive affect you give to the techniques you describe, so they are judged on their own merits.[116]

We tend to conform to the beliefs of those around us, and are especially inclined to avoid being the first dissenter, for social reasons. Being the first dissenter is thus a valuable service.[117] It can be correct if you do not believe you have any special advantage to believe that the majority opinion is more likely to be the true one, but it remains important to express your concerns . Doing so is generally just as socially discouraged as outright disagreement.[118]

Lonely dissent is often just a role people play in defined patterns. When it is real, it requires bearing the incomprehension of the people around you and discussing ideas that are not forbidden but outside bounds which aren't even thought about. Doing this without a single other person is terrifying. Being different for its own sake is a bias like any other.[119]

---

[110] http://lesswrong.com/lw/ls/when_none_dare_urge_restraint/
[111] http://lesswrong.com/lw/lt/the_robbers_cave_experiment/
[112] http://lesswrong.com/lw/lv/every_cause_wants_to_be_a_cult/
[113] http://lesswrong.com/lw/lz/guardians_of_the_truth/
[114] http://lesswrong.com/lw/m0/guardians_of_the_gene_pool/
[115] http://lesswrong.com/lw/m1/guardians_of_ayn_rand/
[116] http://lesswrong.com/lw/m4/two_cult_koans/
[117] http://lesswrong.com/lw/m9/aschs_conformity_experiment/
[118] http://lesswrong.com/lw/ma/on_expressing_your_concerns/
[119] http://lesswrong.com/lw/mb/lonely_dissent/

Cults vary from sincere but deluded and expensive groups, to “love bombing”, sleep deprivation, induced fatigue, distant communes, and daily meetings to confess impure thoughts. Lists of cult characteristics include things which describe other organisations, like political parties and corporations. The true defining aspect is the affective death spiral, which should be fought in any group, and judged independently of how weird the group is in other respects.[120]

## How To Actually Change Your Mind K: Letting Go

If we only admit small, local errors, we only make small, local improvements. Big improvements require admitting big errors. Rather than grudgingly admitting the smallest errors possible, be willing to consider that you may have made fundamental mistakes.[121]

Reinterpreting your mistakes to make it so that you were right ‘deep down’, or morally right, or half-right, avoids the opportunity to see large errors in the route you are on and adjust.[122] Being ready to admit you lost lets you avoid turning small mistakes into bigger ones.[123]

A doubt exists to potentially destroy a particular belief, on the basis of some specific justification. A doubt that fails to either be destroyed or destroy its belief may as well not have existed at all. Wearing doubts as attire does not make you more rational.[124]

You can face reality. *What is true is already so. Owning up to it doesn’t make it any worse.*[125]

Criticising yourself from a sense of duty leaves you wanting to have investigated, not wanting to investigate. This leads to motivated stopping. There is no substitute for genuine curiosity, so attempt to cultivate it. Conservation of expected evidence means any process you think may confirm your beliefs you must also think may disconfirm them. If you do not, ask whether you are looking at only the strong points of your belief.[126]

The laws governing evidence and belief are not social, but aspects of reality. They are not created by rationalists, but merely guessed at. No one can excuse you from them, any more than they may excuse you from the laws of gravity, regardless of how unfair they are in either case.[127]

When you have a cherished belief, ask yourself what you would do, assuming that it was false. Visualise the world in which it is false, without challenging that assumption. Answering this

---

[120] http://lesswrong.com/lw/md/cultish_countercultishness/
[121] http://lesswrong.com/lw/i9/the_importance_of_saying_oops/
[122] http://lesswrong.com/lw/j8/the_crackpot_offer/
[123] http://lesswrong.com/lw/gx/just_lose_hope_already/
[124] http://lesswrong.com/lw/ib/the_proper_use_of_doubt/
[125] http://lesswrong.com/lw/id/you_can_face_reality/
[126] http://lesswrong.com/lw/jz/the_meditation_on_curiosity/
[127] http://lesswrong.com/lw/k1/no_one_can_exempt_you_from_rationalitys_laws/

grants yourself a *line of retreat*- a calm, tolerable path forward- enabling you to consider the question.[128]

When you are invested heavily and emotionally in a long-lived belief which is surrounded by arguments and refutations, it can be desirable to attempt to instigate a real crisis of faith about it, one that could go either way, as it will take more than an ordinary effort to displace if false.[129] [130]

## The Machine In The Ghost L: The Simple Math of Evolution

There are things which look purposeful in nature, which people historically treated as evidence of a designer. If you look at them without cherrypicking, you find parts which appear to be working at odds with other parts, inconsistent with the purposefulness you'd expect from a single designer. Similarly, you find a lot of the purposefulness seems cruel, inconsistent with benevolent design.

If evolution were able to explain anything, it would be useless. Evolution is consistent only with the kind of purposefulness which propagates a gene, with no filtering for kindness or any other kind of purposefulness. This is the kind of alien purposefulness we observe in nature.[131]

Evolution works incrementally.[132] Evolution is slow; a mutation multiplying the expected number of children by 1.03 has a 6% chance of reaching fixation, and takes an average of 768 generations to reach universality within a population of 100,000. The general formulae are 2 **s** for the chance of fixation, and 2 ln(**N**) / **s** for number of generations, where **N** is the population size, and **s** is the multiplier minus 1. Complex mutations take a very long time, as each step must reach fixation.[133]

Price's Equation is a very general equation stating that the change in average characteristic is equal to the covariance of the characteristic and relative fitness. It operates only to the extent that characteristics are heritable across the generations. If characteristics aren't passed down more than a few generations, you will only ever observe a few generations' worth of selective pressure.

This means corporations do not significantly benefit from evolution. Similar for nanodevices with cryptographically protected replication instructions, as few changes would have high covariance.[134]

---

[128] http://lesswrong.com/lw/o4/leave_a_line_of_retreat/
[129] http://lesswrong.com/lw/ur/crisis_of_faith/
[130] http://lesswrong.com/lw/us/the_ritual/
[131] http://lesswrong.com/lw/kr/an_alien_god/
[132] http://lesswrong.com/lw/ks/the_wonder_of_evolution/
[133] http://lesswrong.com/lw/kt/evolutions_are_stupid_but_work_anyway/
[134] http://lesswrong.com/lw/l6/no_evolutions_for_corporations_or_nanodevices/

Selection being concerned only with competition between genes means genes that are better for the species can be outcompeted. Successful genes could make all descendants male, recursively, exist only to copy themselves, or cause the bystander effect. It is possible to evolve to extinction.[135]

Group selection overriding individual selection is generally mathematically implausible and was used to rationalise beliefs that outcomes would be what was better-for-the-species.[136]

Humans are very good at arguing that almost any optimisation criteria suggests almost any policy. Evolution is one of the few cases where we can examine what actually optimising for specific criteria with no rationalisation or bias would look like, in order to understand what that looks like.[137]

We don't consciously have the deliberate goal of optimising for our genes' genetic fitness; it was not genetically fit for that goal to be encoded in us. We are *adaptation-executors, not fitness maximisers*.[138][139] We want to optimise for other things.[140]

Our psychological adaptations are tuned for success in the evolutionary environment.[141] The modern world contains things that match our desires more strongly than anything in the evolutionary environment. We call these *superstimuli*, and they may cause perverse behaviour.[142]

## The Machine In The Ghost M: Fragile Purposes

When observing an intelligent process, you can be certain about the expected end state while being uncertain about intermediary steps. This is because intelligence is an optimisation process.[143] We normally model intelligence by simulating it with our brain, and assume something analogous to our emotional architecture. This doesn't work well for non-human intelligence.[144]

Optimisation processes can find very small targets in large search spaces. Natural selection emerged accidentally, and is slow and stupid. Human brains are much better. Neither optimisation process is able to optimise itself. We could design an AI to do so. If the process did not require exponentially more optimisation power applied for each increase in optimisation

---

[135] http://lesswrong.com/lw/l5/evolving_to_extinction/
[136] http://lesswrong.com/lw/kw/the_tragedy_of_group_selectionism/
[137] http://lesswrong.com/lw/kz/fake_optimization_criteria/
[138] http://lesswrong.com/lw/l0/adaptationexecuters_not_fitnessmaximizers/
[139] http://lesswrong.com/lw/l1/evolutionary_psychology/
[140] http://lesswrong.com/lw/l3/thou_art_godshatter/
[141] http://lesswrong.com/lw/yj/an_especially_elegant_evpsych_experiment/
[142] http://lesswrong.com/lw/h3/superstimuli_and_the_collapse_of_western/
[143] http://lesswrong.com/lw/v8/belief_in_intelligence/
[144] http://lesswrong.com/lw/so/humans_in_funny_suits/

power out, and the initial intelligence was sufficient, optimisation power could rise exponentially over time.[145]

People tend to think of programming computers as if they contain a little ghost which reads and performs abstract instructions. Your instructions define the entirety of the logic performed. If you do not know how to define something in terms you can program, you cannot reference it. Conversely, there is no additional entity capable of deciding to not do what you defined.[146] When we find a confusing gap in our knowledge, we should try to fill it rather than reason around it.[147]

*Terminal values* are ends, *instrumental values* are means.[148] Any generalisations at the macroscopic level will have exceptions; they will be *leaky abstractions*. This extends to instrumental values.[149] We must make any sufficiently powerful and intelligent optimisation process optimise for our terminal values, as optimising for a described instrumental value may powerfully optimise for an easy exception we didn't think of.[150]

*Anthropomorphic optimism* is where we expect non-human intelligent processes, such as natural selection, to choose a strategy that is one a human might choose, because we tend not to bring candidate strategies we know no person wants to the surface, and we're good at rationalization.[151]

Dysfunctional organisations incentivise many actions internally which are detached from any original purpose of the action, and this can be recognised. Civilisation in general does this.[152]

## The Machine In The Ghost N: A Human's Guide To Words

Statements are only entangled with reality if the process generating them made them so.[153]

The logical implications of a given definition of a word are the same in all conceivable universes, and so do not tell us anything about our universe. Correlations between attributes do, but only so far as observations and those correlations are reliable.[154]

If you define a word rigidly in terms of attributes, and then state that something is that word, you assert it has all those attributes. If you then go on to say say it thus has one of those attributes, you are simply repeating that assertion. The word only creates an *illusion of inference*.[155]

---

[145] http://lesswrong.com/lw/rk/optimization_and_the_singularity/
[146] http://lesswrong.com/lw/rf/ghosts_in_the_machine/
[147] http://lesswrong.com/lw/l9/artificial_addition/
[148] http://lesswrong.com/lw/l4/terminal_values_and_instrumental_values/
[149] http://lesswrong.com/lw/lc/leaky_generalizations/
[150] http://lesswrong.com/lw/ld/the_hidden_complexity_of_wishes/
[151] http://lesswrong.com/lw/st/anthropomorphic_optimism/
[152] http://lesswrong.com/lw/le/lost_purposes/
[153] http://lesswrong.com/lw/ne/the_parable_of_the_dagger/
[154] http://lesswrong.com/lw/nf/the_parable_of_hemlock/
[155] http://lesswrong.com/lw/ns/empty_labels/

If assigning a word a definition feels like it argues something, you may be making a hidden assertion of a connotation not in that definition.[156] Alternatively, you may be incorrectly ignoring more direct evidence in favour of correlations between attributes represented by the words.[157]

A *concept* is any rule for classifying things, and creates a category of things. The space of definable concepts is much larger than the space of describable things. We limit ourselves to relatively simple concepts in order to make their definition tractable.[158] Words are labels for concepts.[159]

Efficient communication uses shorter messages for common messages and longer messages for uncommon messages. We use shorter words for more common concepts and longer words for less common concepts.[160] Creating a word defined by a list of attributes permits faster communication if and only if those attributes are correlated. Adding an uncorrelated attribute to a word means it takes more work to communicate accurately using that word than not using it, which will result in inaccurate communication.[161]

We automatically infer that the set of attributes that define a word are well correlated. We shouldn't create definitions where that's wrong.[162] Concepts can be misleading if they group things poorly. Using concepts that are similar to those used by others aids communication.[163] Concepts dividing or excluding things on irrelevant criteria result in people assuming that there's relevant differences correlated to those criteria.[164]

An *intensional definition* is a definition in terms of other words. An *extensional definition* is a definition provided by pointing at examples. The *intension* of a concept is the pattern in your brain that recognises it. The *extension* of a concept is everything matching that pattern. Neither type of definition fully describes its corresponding aspect.

Claiming that a concept with known extension includes a particular attribute 'by definition' hides the assertion that the things in its extension have that attribute. Claiming that a thing falls under a concept 'by definition' often hides the assertion that its attributes are typical of that concept.[165] Not all concept we have, have straightforward intensional definitions. Which concepts usefully divide the world is a question about the world.[166]

---

[156] http://lesswrong.com/lw/ny/sneaking_in_connotations/
[157] http://lesswrong.com/lw/nz/arguing_by_definition/
[158] http://lesswrong.com/lw/o3/superexponential_conceptspace_and_simple_words/
[159] http://lesswrong.com/lw/o9/words_as_mental_paintbrush_handles/
[160] http://lesswrong.com/lw/o1/entropy_and_short_codes/
[161] http://lesswrong.com/lw/o2/mutual_information_and_density_in_thingspace/
[162] http://lesswrong.com/lw/ng/words_as_hidden_inferences/
[163] http://lesswrong.com/lw/nr/the_argument_from_common_usage/
[164] http://lesswrong.com/lw/nx/categorizing_has_consequences/
[165] http://lesswrong.com/lw/nh/extensions_and_intensions/
[166] http://lesswrong.com/lw/o0/where_to_draw_the_boundary/

You can think of any conceivable thing as described by a point in ‘thingspace’, whose dimensions include all possible attributes. Concepts describe clusters in thingspace.[167] These are *similarity clusters*. A dictionary is best thought as a set of hints for matching labels to these clusters.[168] People regard some entities in these clusters as more or less typical of them.[169]

Asking if something ‘is’ in some category is a *disguised query* for whether it should be treated the way things in that category are treated, for some purpose. You may need to know that purpose to answer the question for atypical cases.[170]

You can reduce connections in a neural network design by introducing nodes for categories, then inferring attributes from categories and categories from attributes rather than all attributes from all other attributes.[171] Our brain uses a structure like this. If only some attributes match a category, the way this feels from the inside is like there’s a permanently unresolved question of fact about whether the thing is ‘in’ or not ‘in’ the category, because the ‘node’ is unsettled.[172]

Disputes over definitions are disputes over what cluster a given label points at, but feel like disputes over what properties the things in that cluster have.[173] What intension is associated with what word feels like a fact about the wider world rather than just a fact about human brains.[174]

If you are trying to discuss reality, and you find your meaning for a label differs from another person’s, you should *taboo* that concept and use others to communicate.[175] You can also taboo concepts and try to describe the relevant parts of thingspace directly as an effective way to clarify anticipated experience and notice which aspects of the concepts are relevant.[176]

Our map of the world is necessarily smaller than the world, which means we necessarily must compress distinct things in reality into a single point in our map. From the inside, this feels like we’re observing only one thing, rather than that we’re observing multiple things and compressing them together. Noticing where splitting a category is necessary is a key challenge in reasoning about the world. A good hint is noticing a category with self-contradictory attributes.

---

[167] http://lesswrong.com/lw/nl/the_cluster_structure_of_thingspace/
[168] http://lesswrong.com/lw/nj/similarity_clusters/
[169] http://lesswrong.com/lw/nk/typicality_and_asymmetrical_similarity/
[170] http://lesswrong.com/lw/nm/disguised_queries/
[171] http://lesswrong.com/lw/nn/neural_categories/
[172] http://lesswrong.com/lw/no/how_an_algorithm_feels_from_inside/
[173] http://lesswrong.com/lw/np/disputing_definitions/
[174] http://lesswrong.com/lw/nq/feel_the_meaning/
[175] http://lesswrong.com/lw/nu/taboo_your_words/
[176] http://lesswrong.com/lw/nv/replace_the_symbol_with_the_substance/

[177] Correct statements about different things merged into a single point may be inconsistent with each other; this does not mean part of reality is inconsistent.[178]

Two variables have *mutual information* if they are correlated, and are independent if not. *Conditional independence* is where mutual information is shared between three or more variables, and conditional on one of those variables, the other two become independent. Where we have mutual information between many possible attributes of a thing, we create concepts to represent mutual information between attributes, and then treat the attributes as conditionally independent once we know that something matches that concept, as a simplification.

If there is a great deal of mutual information remaining between attributes after knowing something matches a concept defined using those attributes, this is an error.[179]

Words can be defined wrongly, in many ways.[180]

## Mere Reality O: Lawful Truth

Apparently independent surface-level rules of reality follow from more basic common rules. This means you can't have a consistent world in which some surface-level rules keep working for the same reasons they always worked and others don't work.[181]

The universe almost certainly runs on absolute laws with no exceptions, although we have a much greater degree of uncertainty as to what those laws are. This feels like an unreasonably uncompromising social move to people used to thinking about human or moral laws.[182]

Reality remains uncertain because we don't know the laws, because it isn't feasible to work out the exact consequences of the laws, and we don't know which human in reality we will perceive ourselves as being. Reality is not fundamentally messy; only our perspective on it is.[183]

Bayesian theorems are attractive because they're laws, rather than because Bayesian methods are always the most practical tool.[184] Mutual information is Bayesian evidence; anything which generates better than random beliefs must do so through processing Bayesian evidence.[185]

---

[177] http://lesswrong.com/lw/nw/fallacies_of_compression/
[178] http://lesswrong.com/lw/oc/variable_question_fallacies/
[179] http://lesswrong.com/lw/o8/conditional_independence_and_naive_bayes/
[180] http://lesswrong.com/lw/od/37_ways_that_words_can_be_wrong/
[181] http://lesswrong.com/lw/hq/universal_fire/
[182] http://lesswrong.com/lw/hr/universal_law/
[183] http://lesswrong.com/lw/ms/is_reality_ugly/
[184] http://lesswrong.com/lw/mt/beautiful_probability/
[185] http://lesswrong.com/lw/o7/searching_for_bayesstructure/

A scientist who is not more selective in their beliefs outside the laboratory than a typical person has memorised rules to get by, but lacks understanding of what those rules mean.[186]

No part of a system can violate the first law of thermodynamics, conservation of energy, and so we reject systems claiming to. Liouville's theorem says the space of possible states of a system is conserved; for any part whose state becomes more certain, another part becomes less certain.

The second law of thermodynamics, that total entropy cannot decrease, is a corollary. Maxwell's demon is a hypothetical entity which lets only fast-moving gas molecules through a barrier without generating entropy, decreasing entropy. If you knew the state of the gas for free, you could create one. This means that knowing things about the universe without observing them and generating entropy in the process would be a violation of the second law of thermodynamics.[187]

When people try to justify something without evidence, they often construct theories complicated enough that they can make a mistake and miss it, similar to people designing perpetual motion machines.[188]

## Mere Reality P: Reductionism 101

For some questions, we should, rather than trying to answer or prove them nonsensical, try to identify why we feel a question exists. The result should dissolve that feeling.[189]

A cue that you're dealing with a confused question is when you cannot imagine any observation that answers it.[190] One way forward is to ask "Why do I think <thing>?" rather than "Why <thing>?". The new question will lead you to the entanglement of your beliefs with reality that generated the belief, if it is not confused, and an explanation of your mind otherwise.[191]

The *mind projection fallacy* is treating properties of our perception of a thing as inherent attributes of it.[192] The probability of an event is a property of our perception, not the event.[193] We call something chaotic when we can't predict it, but miss that this is a fact about our ability to predict. This causes us to miss opportunities to improve.[194] Rather than viewing reality as weird, resist getting caught up in incredulity, and let intuition adjust to view reality as normal.[195]

---

[186] http://lesswrong.com/lw/gv/outside_the_laboratory/
[187] http://lesswrong.com/lw/o5/the_second_law_of_thermodynamics_and_engines_of/
[188] http://lesswrong.com/lw/o6/perpetual_motion_beliefs/
[189] http://lesswrong.com/lw/of/dissolving_the_question/
[190] http://lesswrong.com/lw/og/wrong_questions/
[191] http://lesswrong.com/lw/oh/righting_a_wrong_question/
[192] http://lesswrong.com/lw/oi/mind_projection_fallacy/
[193] http://lesswrong.com/lw/oj/probability_is_in_the_mind/
[194] http://lesswrong.com/lw/wb/chaotic_inversion/
[195] http://lesswrong.com/lw/hs/think_like_reality/

Probability assignments are not well modelled as true or false, but as having a level of accuracy. Your beliefs about your own beliefs have different accuracy to those beliefs. Differing beliefs are only differing truths insofar as accurate statements about your own map differ; this is not accurate statements about reality differing between people, because *the map is not the territory*.[196] The concept of a thing is not the same as the thing. If a person thinks a thing is two separate things, described by separate concepts, those concepts may differ despite referring to the same thing.[197]

Reductionism is disbelief in a particular form of the mind projection fallacy. It is useful for us to use different models for different scales of reality, but this is an aspect of what is useful for us, not an aspect of the different scales of reality, and does not mean that they are governed differently.[198]

Explaining and *explaining away* are different. Non-fundamental things still exist. Explaining away something only removes it from the map; it was never in the territory.[199] A thing is only reduced if you know the explanation; knowing one exists only changes literary genre.[200] We can tell human stories about humans. A non-anthropomorphic view of the world helps broader stories.[201]

## Mere Reality Q: Joy In The Merely Real

You should be able to care about knowable, unmagical things. The alternative is existential ennui, because everything is knowable.[202] Taking joy only in discovering something no one else knows makes joy scarce; instead, find joy in all discoveries.[203]

By placing hope in and celebrating true things, you direct your emotions into reality rather than fiction.[204] If we lived in a world with magic, it would seem as mundane as science. If you can't be excited by reality or put in great effort to change the world here, you wouldn't there.[205]

Many of our abilities, such as 'vibratory telepathy' (speech) and 'psychometric tracing' (writing) would be amazing magical powers if only a few had them. Even more so for the 'Ultimate Power'; possessing a small imperfect echo of the universe, and searching through probability to find paths to a desired future. We shouldn't think less of them for commonality.[206]

---

196 http://lesswrong.com/lw/om/qualitatively_confused/
197 http://lesswrong.com/lw/ok/the_quotation_is_not_the_referent/
198 http://lesswrong.com/lw/on/reductionism/
199 http://lesswrong.com/lw/oo/explaining_vs_explaining_away/
200 http://lesswrong.com/lw/op/fake_reductionism/
201 http://lesswrong.com/lw/oq/savanna_poets/
202 http://lesswrong.com/lw/or/joy_in_the_merely_real/
203 http://lesswrong.com/lw/os/joy_in_discovery/
204 http://lesswrong.com/lw/ot/bind_yourself_to_reality/
205 http://lesswrong.com/lw/ou/if_you_demand_magic_magic_wont_help/
206 http://lesswrong.com/lw/ve/mundane_magic/

Settled science is as beautiful as new science. Textbooks will offer you careful explanations, examples, test problems, and likely true information. Pop science articles offer wrong explanations of results the author likely didn't understand, and have a high chance of not replicating. You cannot understand the world if you only read science reporting.[207][208]

Irreligious attempts to imitate religious trappings and hymns always suck. However, a sense of awe is not exclusive to religion. There are things which would have been a good idea even if religion had never existed to imitate that can be awe-inspiring, such as space shuttle launches. For those things, the awe remains when they are mundane and explained.[209]

Things become more desirable as they become less attainable; this is *scarcity*. Similarly, forbidden information appears more important. When something is attained it stops being scarce, leading to frustration.[210] If Science was secret, it would become fascinating.[211][212]

Mysteriousness, faith, unique incommunicability, separation of domains, and experientialism shield from criticism, and declare the mundane boring. We shouldn't have them.[213]

## Mere Reality R: Physicalism 201

Concepts such as 'your hand', describe the same part of the world as lower level concepts, such as 'your palm and fingers'. They do not vary independently, but still 'exist'.[214] Concepts such as 'heat' and 'motion', can also refer to the same thing, even if you can imagine a world where they refer to separate things.[215] Concepts note only that a cluster exists, and do not define it exactly.[216]

Understanding how higher-level things such as 'anger' are created by lower-level things requires discovering the explanation, not just assertion.[217] Rationality is not social rules; rationality is how our brain works.[218] Reality is that which sometimes violates expectations and surprises you.[219]

---

[207] http://lesswrong.com/lw/ow/the_beauty_of_settled_science/
[208] http://lesswrong.com/lw/ox/amazing_breakthrough_day_april_1st/
[209] http://lesswrong.com/lw/oy/is_humanism_a_religionsubstitute/
[210] http://lesswrong.com/lw/oz/scarcity/
[211] http://lesswrong.com/lw/p0/to_spread_science_keep_it_secret/
[212] http://lesswrong.com/lw/p1/initiation_ceremony/
[213] http://lesswrong.com/lw/57/the_sacred_mundane/
[214] http://lesswrong.com/lw/p2/hand_vs_fingers/
[215] http://lesswrong.com/lw/p4/heat_vs_motion/
[216] http://lesswrong.com/lw/p6/reductive_reference/
[217] http://lesswrong.com/lw/p3/angry_atoms/
[218] http://lesswrong.com/lw/k2/a_priori/
[219] http://lesswrong.com/lw/p6/reductive_reference/

The brain is a complex organ made of neurons.[220] Before we realised that thinking involved a complex organ, Animism was a reasonable error.[221] A proposed entity is supernatural if it is irreducibly complex. Because our brains are reducible, no set of expectations can require irreducible complexity, but some expectations make irreducibility more likely than others.[222][223]

A *zombie*, in the philosophical sense, is a hypothetical being which looks and behaves exactly like a human, including talking about being conscious, but is not conscious. It is alleged that if it is a coherent hypothetical, consciousness must be extra-physical. It is not coherent if 'process which causes talking about consciousness' and 'consciousness' refer to the same part of the world. We should believe they do, because the alternative is more complex.[224][225][226] It is correct to believe in unobservable things if and only if the most succinct model of reality predicts them.[227]

The *generalised anti-zombie principle* is that any change we shouldn't expect to change the reasons we talk about consciousness is one we should expect to leave us still conscious.[228] Conceivably, one could replace a human with a giant look-up table (GLUT) which would seem to violate this principle, but the process which selected the GLUT to use would need to have been conscious and make all the same decision-making choices as you in doing so.[229]

## Mere Reality S: Quantum Physics and Many Worlds

*(This sequence is controversial; mean probability assigned to MWI was 56.5% in the 2011 survey)*

Quantum mechanics is not intuitive; this is a flaw in intuition.[230]

Reality is comprised of *configurations* with complex-valued *amplitudes*, and rules for calculating amplitude flows into other configurations. We cannot measure amplitudes directly, only the ratio of absolute squares of some configurations.[231] You sum all amplitude flows into a configuration to get its amplitude. Amplitude flows that put the same types of particle in the same places flow into the same configuration, even if the particles came from different places. Which configurations are the same is observable fact. If amplitude flows have opposite sign, they can

---

[220] http://lesswrong.com/lw/p5/brain_breakthrough_its_made_of_neurons/
[221] http://lesswrong.com/lw/t5/when_anthropomorphism_became_stupid/
[222] http://lesswrong.com/lw/tv/excluding_the_supernatural/
[223] http://lesswrong.com/lw/tw/psychic_powers/
[224] http://lesswrong.com/lw/p7/zombies_zombies/
[225] http://lesswrong.com/lw/p8/zombie_responses/
[226] http://lesswrong.com/lw/pn/zombies_the_movie/
[227] http://lesswrong.com/lw/pb/belief_in_the_implied_invisible/
[228] http://lesswrong.com/lw/p9/the_generalized_antizombie_principle/
[229] http://lesswrong.com/lw/pa/gazp_vs_glut/
[230] http://lesswrong.com/lw/pc/quantum_explanations/
[231] http://lesswrong.com/lw/pd/configurations_and_amplitude/

cancel out to zero. If either flow had been absent, the configuration would have had non-zero amplitude.[232]

A configuration is defined by all particles. If amplitude flows alter a particle's state, then they cannot flow into the same configuration as amplitude flows which do not alter it. Thus, measuring amplitude flows stops them from flowing to the same configurations.[233]

*Collapse* theories propose that at some point before a measurement reaches a human brain, there is a waveform collapse leaving only one random configuration with non-zero amplitude, discarding other amplitude flows. *Many Worlds* proposes that this doesn't happen; configurations where we observe and don't observe a measurement both exist with non-zero amplitude, too different from each other for their amplitude flows to flow into common configurations; we have *macroscopic decoherence*. Collapse would be very different to other physics.[234] Living in multiple worlds is the same as living in one; we shouldn't be unsettled by it.[235]

Decoherence is simpler[236], while making the same predictions.[237] *Privileging the hypothesis* is selecting an unlikely hypothesis for attention, causing confirmation bias. Historical accident has privileged collapse theories,[238][239][240] because people didn't think of themselves as made of particles.[241][242] Declaring equations to be meaningless is wrong; there is something described.[243]

## Mere Reality T: Science and Rationality

Science is supposed to replace theories when experiments falsify them in favour of new theories, and is uninterested in simpler theories making the same predictions. This leads to different results than application of probability theory.[244] Science is this way because it doubts that flawed humans debating elegance will reach truth if not forced to experiment. Science distrusts your rationality.[245]

Science doesn't help you get answers to questions that are not testable in the present day. It is incorrect to dismiss theories answering those questions because they're scientifically unproven.

---

[232] http://lesswrong.com/lw/pe/joint_configurations/
[233] http://lesswrong.com/lw/pf/distinct_configurations/
[234] http://lesswrong.com/lw/q6/collapse_postulates/
[235] http://lesswrong.com/lw/qz/living_in_many_worlds/
[236] http://lesswrong.com/lw/q3/decoherence_is_simple/
[237] http://lesswrong.com/lw/q4/decoherence_is_falsifiable_and_testable/
[238] http://lesswrong.com/lw/19m/privileging_the_hypothesis/
[239] http://lesswrong.com/lw/q7/if_manyworlds_had_come_first/
[240] http://lesswrong.com/lw/q8/many_worlds_one_best_guess/
[241] http://lesswrong.com/lw/pg/where_philosophy_meets_science/
[242] http://lesswrong.com/lw/r0/thou_art_physics/
[243] http://lesswrong.com/lw/q5/quantum_nonrealism/
[244] http://lesswrong.com/lw/qa/the_dilemma_science_or_bayes/
[245] http://lesswrong.com/lw/qb/science_doesnt_trust_your_rationality/

You must try to use your reason.[246] Science does not judge your choice of hypothesis, and only requires you react to overwhelming evidence. It accepts slow, generational progress. You must have a private epistemic standard higher than the social one, or else you will waste a lot of time.[247]

It is a flaw that the teaching of Science doesn't practice resolving confused ideas,[248] probability theory, awareness of the need for causal entanglement of belief with reality, or rationality more broadly.[249] Teaching probability theory alone would not correct this.[250]

There is nothing that guarantees that you are not a fool, not even Science, not even trying to use probability theory. You don't know your own biases, why the universe is simple enough to understand, what your priors are, or why they work. The formal math is intractable. To start as a rationalist requires losing your trust that following any prescribed pattern will keep you safe.[251]

The bulk of work in progressing knowledge is in elevating the right hypotheses to attention, a process Science depends on but does not specify, relying on normal reasoning.[252] Einstein did this well. Most will fail, but it remains valuable to practice.[253] Geniuses are not separate from humanity; with grit and the right choice of problem and approach, not all but many have potential.[254][255]

We do not use the evidence of sensory data anywhere near optimally.[256] Possible minds can be extremely smarter than humans. Basing your ideals on hypothetical extremely intelligent minds, rather than merely the best humans so far, helps you not shy away from trying to exceed them.[257]

## Mere Goodness U: Fake Preferences

Human desires include preferences for how the world is, not just preferences for how they think the world is or how happy they are.[258] People who claim their preferences reduce down to a single principle have some other process by which they choose what they want, and then find a rationalisation for how what they want is justified by that principle.[259] Simple utility functions fail

---

[246] http://lesswrong.com/lw/qc/when_science_cant_help/
[247] http://lesswrong.com/lw/qd/science_isnt_strict_enough/
[248] http://lesswrong.com/lw/q9/the_failures_of_eld_science/
[249] http://lesswrong.com/lw/qe/do_scientists_already_know_this_stuff/
[250] http://lesswrong.com/lw/qg/changing_the_definition_of_science/
[251] http://lesswrong.com/lw/qf/no_safe_defense_not_even_science/
[252] http://lesswrong.com/lw/qi/faster_than_science/
[253] http://lesswrong.com/lw/qj/einsteins_speed/
[254] http://lesswrong.com/lw/qs/einsteins_superpowers/
[255] http://lesswrong.com/lw/qt/class_project/
[256] http://lesswrong.com/lw/qk/that_alien_message/
[257] http://lesswrong.com/lw/ql/my_childhood_role_model/
[258] http://lesswrong.com/lw/lb/not_for_the_sake_of_happiness_alone/
[259] http://lesswrong.com/lw/kx/fake_selfishness/

to compress our values, and we suffer from anthropomorphic optimism about what they suggest.[260]

People who fear that humans would lack morality without an external threat, regard this as bad rather than liberating. This means they like morality, and aren't just forced to abide by it.[261]

The *detached lever fallacy* is the assumption that actions that trigger behaviour from one entity will trigger it from another, without any reason to think the mechanics governing the reaction are present in the second. The actions that make a human compassionate will not make a non-human AI so.[262] AI design is reducing the mental to the non-mental. Models of an intelligence which can't predict what it will do other than by analogy to a human are incomplete.[263] The space of possible minds is extremely large. Resist the temptation to generalise over all of mind design space.[264]

## Mere Goodness V: Value Theory

Justifying any belief leads to infinite regress. Rather than accepting any assumption, we should reflect on our mind's trustworthiness using our current mind as best we can, and accept that.[265] Approach such questions from the standpoint of whether we should want ourselves or an AI using similar principles to change how they choose beliefs. We should focus on improvement, not justification, and expect to change our minds. Don't exalt consistency in itself, but effectiveness. Separate asking "why" an approach works from whether it "does". We should reason about our own mind the way we do about the rest of the world, and use all available information.[266]

There are no arguments compelling to all possible minds. For any system processing information, there is a system with inverted output which makes the opposite conclusion. This applies to moral conclusions, and regardless of the intelligence of the system.[267][268] A mind must have a process that adds beliefs, and a process that acts, or no argument can convince it to believe or act.[269]

Some properties can be either thought of as as taking two parameters and giving a result, or as a space of one-parameter functions, with different people using different ones. For example, 'attractiveness(admirer, admired) -> result' vs 'attractiveness_*1...9999*(admired) -> result'.

---

[260] http://lesswrong.com/lw/lq/fake_utility_functions/
[261] http://lesswrong.com/lw/ky/fake_morality/
[262] http://lesswrong.com/lw/sp/detached_lever_fallacy/
[263] http://lesswrong.com/lw/tf/dreams_of_ai_design/
[264] http://lesswrong.com/lw/rm/the_design_space_of_mindsingeneral/
[265] http://lesswrong.com/lw/s0/where_recursive_justification_hits_bottom/
[266] http://lesswrong.com/lw/s2/my_kind_of_reflection/
[267] http://lesswrong.com/lw/rn/no_universally_compelling_arguments/
[268] http://lesswrong.com/lw/sy/sorting_pebbles_into_correct_heaps/
[269] http://lesswrong.com/lw/rs/created_already_in_motion/

*Currying* specifies that a two parameter function is equivalent to a one parameter function returning another function, and unifies these. For example, 'attractiveness(admirer) -> attractiveness_712(admired) -> result'. This reflects the ability to judge a measure independently of the user, but also that the measure used is variable.[270]

If your moral framework is shown to be invalid, you can still choose to act morally anyway.[271] It's important to have a line of retreat to be able to seriously review your metaethics.[272] You must start from a willingness to evaluate in terms of your moral intuition in order to find valid metaethics.[273] What we consider to be right grows out of a starting point. To get a system that specifies what is right requires it fit that starting point, which we cannot define fully.[274] Concepts that we develop to describe good behaviour are very complex. Depictions of them have many possible concepts that fit them, and an algorithm would pick the wrong one.You cannot fix a powerful optimisation process optimising for the wrong thing with patches.[275] Value is fragile; optimising for the wrong values creates a dull future.[276] Our complicated values are the gift that we give to tomorrow.[277]

The *prisoner's dilemma* is a hypothetical in which two people can both either cooperate (C) or defect (D), and each one prefers (D, C) > (C, C) > (D, D) > (C, D). The typical example involves two totally selfish prisoners, but humans can't imagine this. A better example would have the first entity as humans trying to save billions, vs an entity trying to maximise numbers of paperclips.[278]

We understand others by simulating them with our brains, which creates empathy. It was evolutionarily useful to develop sympathy. An AI wouldn't use either approach, an alien might.[279]

A world with no difficulty would be boring, We prefer real goals to fake ones. We need goals which we prefer working on to having finished, or which have no end state.[280] A utopia with no problems has no stories. Pain can be more intense than pleasure. Pleasure that scaled like pain would trap us. We can be rid of pain that breaks or grinds down people, and pointless sorrow, and keep what we value. Whether we will get rid of pain entirely someday, EY does not know.[281]

---

[270] http://lesswrong.com/lw/ro/2place_and_1place_words/
[271] http://lesswrong.com/lw/rq/what_would_you_do_without_morality/
[272] http://lesswrong.com/lw/sk/changing_your_metaethics/
[273] http://lesswrong.com/lw/sb/could_anything_be_right/
[274] http://lesswrong.com/lw/sw/morality_as_fixed_computation/
[275] http://lesswrong.com/lw/td/magical_categories/
[276] http://lesswrong.com/lw/y3/value_is_fragile/
[277] http://lesswrong.com/lw/sa/the_gift_we_give_to_tomorrow/
[278] http://lesswrong.com/lw/tn/the_true_prisoners_dilemma/
[279] http://lesswrong.com/lw/xs/sympathetic_minds/
[280] http://lesswrong.com/lw/ww/high_challenge/
[281] http://lesswrong.com/lw/xi/serious_stories/

# Mere Goodness W: Quantified Humanism

*Scope insensitivity* is ignoring the number of people or animals or area affected, the *scope*, when deciding how important an action is. Groups were asked how much they would pay to save 2000 / 20000 / 200000 migrating birds from drowning in oil ponds, and answered $80, $78, and $88. We visualise a single bird, react emotionally, and cannot visualise scope. To be an effective altruist, we must evaluate the numbers.[282] Saving one life feels as good as many, but is not as good. We do not treat saving lives as a satisficed virtue, such that once you've saved one you ignore others.[283]

The *certainty effect* is a bias where going from 99% chance to near 100% chance of getting what we want is valued more than going from, say, 33% to 34%. This causes the *allais paradox*, where we prefer a fixed prize over a 33/34 chance of a bigger prize, but prefer a 33% chance of a larger prize to a 34% chance of a smaller prize. This cannot be explained by non-linear marginal utility of money, permits extracting money from you, and shows a failure of intuition to steer reality.[284][285]

A certain loss feels worse than an uncertain one. By changing the point of comparison so the certain outcome is a loss rather than a gain, you reverse intuition. You must multiply out costs and benefits, or you will fail at directing reality. This reduces nice feelings, but they are not the point.[286]

Intuition is what morality is built on, but we must pursue reflective intuitions or we won't accomplish anything due to circular preferences.[287] Making up probabilities can trick you into thinking they're more grounded than they are, and override working intuitions.[288]

*Ends don't justify the means among humans*. We run on *corrupted hardware*; we rationalise using bad means, past the point that benefits us, let alone anyone else. Otherwise we wouldn't have developed *ethical injunctions*. Follow them as a higher-level consequentialist strategy.[289][290]

To pursue rationality effectively, you must have a higher goal that it serves.[291] *Newcomb's problem* is a scenario in which an entity that can predict you perfectly offers two boxes, and says that box A contains $1000, and box B contains $1,000,000 if and only if they predicted you

---

[282] http://lesswrong.com/lw/hw/scope_insensitivity/
[283] http://lesswrong.com/lw/hx/one_life_against_the_world/
[284] http://lesswrong.com/lw/my/the_allais_paradox/
[285] http://lesswrong.com/lw/mz/zut_allais/
[286] https://wiki.lesswrong.com/wiki/Feeling_Moral
[287] http://lesswrong.com/lw/n9/the_intuitions_behind_utilitarianism/
[288] http://lesswrong.com/lw/sg/when_not_to_use_probabilities/
[289] http://lesswrong.com/lw/uv/ends_dont_justify_means_among_humans/
[290] http://lesswrong.com/lw/v1/ethical_injunctions/
[291] http://lesswrong.com/lw/nb/something_to_protect/

would only take box B. Traditional causal decision theory says you should take both boxes, as the money is either already in the box or not. Rationally, you should take only box B. Doing so makes you win more, and *rationality is about winning*, not about reasonableness or any particular ritual of thought.[292]

## Becoming Stronger X: Yudkowsky's Coming Of Age

Yudkowsky grew up in an environment which praised experience over intelligence as justification for everything, including religion. This led them to the opposite, an affective death spiral around intelligence as the solution to everything. They thought that being very intelligent meant being very moral. They tended to go too far the other way in reaction to someone else's stupidity.[293]

Because previous definitions of intelligence had been lacking, they thought it could not be defined tidily. This led to avoiding premature answers. They believed the field of AI research was sick; this led to studying cognitive science. Errors which lead to studying more are better errors.[294] They regarded regulation of technology as bad, and this reduced attention to existential risks. When convinced risks existed, rather than reviewing mistakes, they just decided we needed AI first.[295]

They were good at refuting arguments, and felt they were winning the debate on whether intelligence implied morality. They had a rationale for proceeding with their best ideas, without resolving confusion. Reality does not care whether you are using your best ideas. You can't rely on anyone giving you a flawless argument, and you can't work around underlying confusion.[296] [297]

An incongruous thought, coupled with some perfectionism, and viewing less than morally upright interactions as unacceptable, led to investigating seriously. Doing that, regardless of reason, led to pursuing a backup plan.[298] That they were pursuing a backup plan gave them a line of retreat for their earlier views, but they only shifted gradually, without acknowledging fundamental errors.[299]

They only saw the error when they realised that a mind was an optimisation process which pumps reality towards outcomes, and you could pump towards any outcomes.[300] They realised that they could have unrefuted arguments, and nature could still kill them if the choice was

---

[292] http://lesswrong.com/lw/nc/newcombs_problem_and_regret_of_rationality/
[293] http://lesswrong.com/lw/ty/my_childhood_death_spiral/
[294] http://lesswrong.com/lw/tz/my_best_and_worst_mistake/
[295] http://lesswrong.com/lw/u0/raised_in_technophilia/
[296] http://lesswrong.com/lw/u1/a_prodigy_of_refutation/
[297] http://lesswrong.com/lw/u2/the_sheer_folly_of_callow_youth/
[298] http://lesswrong.com/lw/u7/that_tiny_note_of_discord/
[299] http://lesswrong.com/lw/u8/fighting_a_rearguard_action_against_the_truth/
[300] http://lesswrong.com/lw/u9/my_naturalistic_awakening/

wrong. Their trust in following patterns broke, and they began studying rationality.[301] We all need to lose our assumption of fairness.[302] They realised that an idea seeming very good didn't permit being sure; it needed to be provably equivalent to any correct alternative, like Bayesian probability.[303]

They recognise that there are people more formidable than them, and hope that their writings might find a younger one of them who can then exceed them.[304]

## Becoming Stronger Y: Challenging The Difficult

Wanting to become stronger means reacting to flaws by doing what you can to repair them rather than with resignation. Do not ritualistically confess your flaws unless you include what you intend to do about them.[305] If you are ashamed of wanting to do better than others, you will not make a real effort to seek higher targets. You should always reach higher, without shame.[306]

The difference between saying that you are going to *do something*, and that you are going to *try to do something*, is that the latter makes you satisfied with a plan, rather than with success, and allows the part where the plan has to maximise your odds of success to get lost. Don't try your best; either win or fail.[307] People don't make genuine efforts to win even for five minutes.[308]

A desperate effort is a level above wanting to become stronger, where you try as though your life were at stake. And there is a step above that, an extraordinary effort; it requires being willing to go outside of a comfortable routine, tackle difficulties you don't have a mental routine for, and bypass usual patterns, in order to achieve an outcome that is not the default that you care greatly about. It is riskier than even a desperate effort.[309]

A problem being impossible sometimes only means that when we query our brain for a strategy, we can't think of one. This is not the same as being proven to be impossible. Genuine effort over years can find routes forward. Reality can uncaringly demand the impossible. We should resist our urge to find rationalisations for why the problem doesn't matter,[310] and sometimes we should *shut up and do the impossible*; take success at the impossible as our goal and accept nothing less.[311]

---

[301] http://lesswrong.com/lw/ue/the_magnitude_of_his_own_folly/
[302] http://lesswrong.com/lw/uk/beyond_the_reach_of_god/
[303] http://lesswrong.com/lw/ul/my_bayesian_enlightenment/
[304] http://lesswrong.com/lw/ua/the_level_above_mine/
[305] http://lesswrong.com/lw/h8/tsuyoku_naritai_i_want_to_become_stronger/
[306] http://lesswrong.com/lw/h9/tsuyoku_vs_the_egalitarian_instinct/
[307] http://lesswrong.com/lw/uh/trying_to_try/
[308] http://lesswrong.com/lw/ui/use_the_try_harder_luke/
[309] http://lesswrong.com/lw/uo/make_an_extraordinary_effort/
[310] http://lesswrong.com/lw/un/on_doing_the_impossible/
[311] http://lesswrong.com/lw/up/shut_up_and_do_the_impossible/

We need to ask ourselves what we want, what it will require to accomplish, and set out to do it with what we know.[312]

## Becoming Stronger Z: The Craft and the Community

The prevalence of religion, even in scientific circles, warns us that the baseline grasp of rationality is very low. Arguing against religion specifically fails to solve the underlying problem. We should also be trying to *raise the sanity waterline*.[313]

A reason that people don't want to learn more about rationality is that they don't see people who know about it as happier or more successful. A large part of this is that even the people who know a lot about it still know very little, compared to experts in other fields; we have not systematised it as a field of study, subject to large-scale investment and experimentation. One reason for this is that traditional rationalists/skeptics do not see lack of visible formidability and say that we must be doing something wrong. We treat it as a mere hobby horse.[314] It can take more than an incremental step in the direction of rationality to get an incremental increase in winning.[315]

Martial arts dojos suffer from *epistemic viciousness*; a treatment of the master as sacred, exaltation of historic knowledge over discovery, a lack of data, and a pretense that lack of data isn't a real problem. Hypothetical rationality dojos risk the same problems.[316] If an air of authority can substitute for evidence, traditions can proliferate and wield influence without evidence.[317]

Verification methods can be stratified into three levels. *Reputational* verification is the basic practice of trying to ground reputations in some real world or competitive performance. *Experimental* verification is randomised, replicable testing, although this can involve very simple measures that are only correlated with the variables of interest. *Organisational* verification is that which, when everyone knows the process, is resistant enough to gaming to continue working.[318]

Groups which do not concern themselves with rationality can praise agreement, encourage the less agreeing to leave, and enter an affective death spiral, which binds them all together and makes them cooperate. Typical rationalist groups do not cooperate; they speak and applaud disagreement but not agreement. If you are outperformed by irrational groups, then you are not rational, because rationality is about winning. Actual rationality should involve being better at coordinating, and we should work out how to be. Being half a rationalist is dangerous.[319][320] Until

---

[312] http://lesswrong.com/lw/cl/final_words/
[313] http://lesswrong.com/lw/1e/raising_the_sanity_waterline/
[314] http://lesswrong.com/lw/2c/a_sense_that_more_is_possible/
[315] http://lesswrong.com/lw/7k/incremental_progress_and_the_valley/
[316] http://lesswrong.com/lw/2i/epistemic_viciousness/
[317] http://lesswrong.com/lw/2j/schools_proliferating_without_evidence/
[318] http://lesswrong.com/lw/2s/3_levels_of_rationality_verification/
[319] http://lesswrong.com/lw/3h/why_our_kind_cant_cooperate/

atheist groups can outperform religious groups at mobilisation and output, any increase in atheism is a hollow victory.[321] We need new models of community to replace the old, with new goals.[322]

Do not punish people for being more patient than you; you should *tolerate tolerance*.[323] We incentivise groups to improve by rejecting joining them if they don't meet our standards. The non-conformist crowd tends to ask way too much. If joining a project is good, you should do it if the problems are not too distracting, or if you could fix the problems. If you don't see a problem as worth putting in the time to fix, it is not worth avoiding a group for. If we want to get anything done, we need to move in the direction of joining groups and staying in them.[324]

Many causes benefit from the spread of rationality. We should not think of other good causes as in competition for a limited pool of reasonable thinkers, but instead cooperate with them to increase the number of reasonable thinkers. We should think of ourselves as all part of one common project of human progress.[325] We are very bad at coordinating to fulfil aligned preferences of individuals. Large flows of money tend to be controlled by the incentives of organisations.[326]

Donating time is inefficient compared to donating money. Allocating money is how we allocate resources. *Money is the unit of caring*. If you'll never spend it, you don't care.[327] We enjoy having done kind things, but the things that bring us enjoyment often do much less good than calculated effort, and enjoyment and social status can be had much cheaper when you don't try to achieve them through your giving. Get enjoyment, status, and results separately; *purchase fuzzies and utilons separately*.[328]

The *bystander effect* is a bias in which a group is less likely to react to an emergency than a single individual.[329] This applies to problems encountered over the Internet, where you are always observing them as part of a group of strangers.[330]

When we write advice, we are not working from universal generalisations, but surface level tricks. This means it validly works for some people but not others. We should *beware other-optimising*, because we are not good at knowing what works for others, and beware

---

320 http://lesswrong.com/lw/5f/bayesians_vs_barbarians/
321 http://lesswrong.com/lw/5t/can_humanism_match_religions_output/
322 http://lesswrong.com/lw/5v/church_vs_taskforce/
323 http://lesswrong.com/lw/42/tolerate_tolerance/
324 http://lesswrong.com/lw/5j/your_price_for_joining/
325 http://lesswrong.com/lw/66/rationality_common_interest_of_many_causes/
326 http://lesswrong.com/lw/64/helpless_individuals/
327 http://lesswrong.com/lw/65/money_the_unit_of_caring/
328 http://lesswrong.com/lw/6z/purchase_fuzzies_and_utilons_separately/
329 http://lesswrong.com/lw/9j/bystander_apathy/
330 http://lesswrong.com/lw/9m/collective_apathy_and_the_internet/

assuming that other people are simply not trying what worked for us.[331] Practical advice based on established theories tends to be more generalisable.[332]

The danger of underconfidence is missing opportunities and not making a genuine effort. Sticking to things you always win at is a way smart people become stupid. You should seriously try to win, but aim for challenges you might lose at. When considering a habit of thought, ask whether it makes you stronger or weaker.[333]

There is more absent than present in these writings. Defeating akrasia and coordinating groups are particular absences. But, hopefully, there is enough to overcome the barriers to getting started in the matter of rationality without immediately going terribly wrong. The hope is that this art of answering confused questions will be enough to *go and complete the rest*. This will require drawing on many sources, and require having some specific motivating goal. *Go forth and create the art*, and return to tell others what you learned.[334]

## And A Few Third-Party Sequences and Primers

Yvain has a primer to game theory. Lukeprog has a sequence on scientifically-backed advice for winning at life, to the extent to which it is available. Orthnonormal has a primer on decision theory and the motivation for discussing alternative decision theories, and their implications, such as acausal trade. These three areas were popular topics for further discussion on Less Wrong.

[331] http://lesswrong.com/lw/9v/beware_of_otheroptimizing/
[332] http://lesswrong.com/lw/d4/practical_advice_backed_by_deep_theories/
[333] http://lesswrong.com/lw/c3/the_sin_of_underconfidence/
[334] http://lesswrong.com/lw/c4/go_forth_and_create_the_art/