Checklist of Rationality Habits

Center For Applied Rationality: Updated July 2012

This checklist is meant for your personal use so you can have a wish-list of rationality habits, and so that you can see if you're acquiring good habits over the next year—we're not using it to decide how rational you are at the start of the program.

1. Reacting to evidence / surprises / arguments you haven't heard before; flagging beliefs for examination.

a) When I see something odd - something that doesn't fit with what I'd ordinarily expect, given my other beliefs - I successfully notice, promote it to conscious attention and think "I notice that I am confused" or some equivalent thereof. (*Example: You think that your flight is scheduled to depart on Thursday. On Tuesday, you get an email from Travelocity advising you to prepare for your flight "tomorrow", which seems wrong. Do you successfully raise this anomaly to the level of conscious attention?* (*Based on the experience of an actual LWer who failed to notice confusion at this point and missed their plane flight.*))

Date of last example:
□ Never
□ Today/yesterday
□ Last week
□ Last month
□ Last year
□ Before the last year

b) When somebody says something that isn't quite clear enough for me to visualize, I notice this and ask for examples. (*Recent example from Eliezer: A mathematics student said they were studying "stacks". I asked for an example of a stack. They said that the integers could form a stack. I asked for an example of something that was not a stack.*) (*Recent example from Anna: Cat said that her boyfriend was very competitive. I asked her for an example of "very competitive." She said that when he's driving and the person next to him revs their engine, he must be the one to leave the intersection first—and when he's the passenger he gets mad at the driver when they don't react similarly.*)

Date of last example:
□ Never
□ Today/yesterday
□ Last week
□ Last month
□ Last year
□ Before the last year

c) I notice when my mind is arguing for a side (instead of evaluating which side to choose), and flag this as an error mode. (*Recent example from Anna: Noticed myself explaining to myself why outsourcing my clothes shopping does make sense, rather than evaluating whether to do it.*)

Date of last example:
□ Never
□ Today/yesterday
□ Last week
□ Last month
□ Last year
□ Before the last year

d) I notice my mind flinching away from a thought; and when I notice, I flag that area as requiring more deliberate exploration. (*Recent example from Anna: I have a failure mode where, when I feel socially uncomfortable, I try to make others feel mistaken so that I will feel less vulnerable. Pulling this thought into words required repeated conscious effort, as my mind kept wanting to just drop the subject.*)

Date of last example:
□ Never
□ Today/yesterday
□ Last week
□ Last month
□ Last year
□ Before the last year

e) I consciously attempt to welcome bad news, or at least not push it away. (*Recent example from Eliezer: At a brainstorming session for future Singularity Summits, one issue raised was that we hadn't really been asking for money at previous ones. My brain was offering resistance, so I applied the "bad news is good news" pattern to rephrase this as, "This point doesn't change the fixed amount of money we raised in past years, so it is good news because it implies that we can fix the strategy and do better next year."*)	Date of last example: □ Never □ Today/yesterday □ Last week □ Last month □ Last year □ Before the last year

2. Questioning and analyzing beliefs (after they come to your attention).

a) I notice when I'm not being curious. (*Recent example from Anna: Whenever someone criticizes me, I usually find myself thinking defensively at first, and have to visualize the world in which the criticism is true, and the world in which it's false, to convince myself that I actually want to know. For example, someone criticized us for providing inadequate prior info on what statistics we'd gather for the Rationality Minicamp; and I had to visualize the consequences of [explaining to myself, internally, why I couldn’t have done any better given everything else I had to do], vs. the possible consequences of [visualizing how it might've been done better, so as to update my action-patterns for next time], to snap my brain out of defensive-mode and into should-we-do-that-differently mode.*)	Date of last example: □ Never □ Today/yesterday □ Last week □ Last month □ Last year □ Before the last year
b) I look for the actual, historical causes of my beliefs, emotions, and habits; and when doing so, I can suppress my mind's search for justifications, or set aside justifications that weren't the actual, historical causes of my thoughts. (*Recent example from Anna: When it turned out that we couldn't rent the Minicamp location I thought I was going to get, I found lots and lots of reasons to blame the person who was supposed to get it; but realized that most of my emotion came from the fear of being blamed myself for a cost overrun.*)	Date of last example: □ Never □ Today/yesterday □ Last week □ Last month □ Last year □ Before the last year
c) I try to think of a concrete example that I can use to follow abstract arguments or proof steps. (*Classic example: Richard Feynman being disturbed that Brazilian physics students didn't know that a "material with an index" meant a material such as water. If someone talks about a proof over all integers, do you try it with the number 17? If your thoughts are circling around your roommate being messy, do you try checking your reasoning against the specifics of a particular occasion when they were messy?*)	Date of last example: □ Never □ Today/yesterday □ Last week □ Last month □ Last year □ Before the last year

d) When I'm trying to distinguish between two (or more) hypotheses using a piece of evidence, I visualize the world where hypothesis #1 holds, and try to consider the prior probability I'd have assigned to the evidence in that world, then visualize the world where hypothesis #2 holds; and see if the evidence seems more likely or more specifically predicted in one world than the other (*Historical example: During the Amanda Knox murder case, after many hours of police interrogation, Amanda Knox turned some cartwheels in her cell. The prosecutor argued that she was celebrating the murder. Would you, confronted with this argument, try to come up with a way to make the same evidence fit her innocence? Or would you first try visualizing an innocent detainee, then a guilty detainee, to ask with what frequency you think such people turn cartwheels during detention, to see if the likelihoods were skewed in one direction or the other?*)	Date of last example: □ Never □ Today/yesterday □ Last week □ Last month □ Last year □ Before the last year
e) I try to consciously assess prior probabilities and compare them to the apparent strength of evidence. (*Recent example from Eliezer: Used it in a conversation about apparent evidence for parapsychology, saying that for this I wanted p < 0.0001, like they use in physics, rather than p < 0.05, before I started paying attention at all.*)	Date of last example: □ Never □ Today/yesterday □ Last week □ Last month □ Last year □ Before the last year
f) When I encounter evidence that's insufficient to make me "change my mind" (substantially change beliefs/policies), but is still more likely to occur in world X than world Y, I try to update my probabilities at least a little. (*Recent example from Anna: Realized I should somewhat update my beliefs about being a good driver after someone else knocked off my side mirror, even though it was legally and probably actually their fault—even so, the accident is still more likely to occur in worlds where my bad-driver parameter is higher.*)	Date of last example: □ Never □ Today/yesterday □ Last week □ Last month □ Last year □ Before the last year

3. Handling inner conflicts; when different parts of you are pulling in different directions, you want different things that seem incompatible; responses to stress.

a) I notice when I and my brain seem to believe different things (a belief-vs-anticipation divergence), and when this happens I pause and ask which of us is right. (*Recent example from Anna: Jumping off the Stratosphere Hotel in Las Vegas in a wire-guided fall. I knew it was safe based on 40,000 data points of people doing it without significant injury, but to persuade my brain I had to visualize 2 times the population of my college jumping off and surviving. Also, my brain sometimes seems much more pessimistic, especially about social things, than I am, and is almost always wrong.*)	Date of last example: □ Never □ Today/yesterday □ Last week □ Last month □ Last year □ Before the last year

b) When facing a difficult decision, I try to reframe it in a way that will reduce, or at least switch around, the biases that might be influencing it. (*Recent example from Anna's brother: Trying to decide whether to move to Silicon Valley and look for a higher-paying programming job, he tried a reframe to avoid the status quo bias: If he was living in Silicon Valley already, would he accept a $70K pay cut to move to Santa Barbara with his college friends?* (*Answer: No.*))	Date of last example: □ Never □ Today/yesterday □ Last week □ Last month □ Last year □ Before the last year
c) When facing a difficult decision, I check which considerations are consequentialist - which considerations are actually about future consequences. (*Recent example from Eliezer: I bought a $1400 mattress in my quest for sleep, over the Internet hence much cheaper than the mattress I tried in the store, but non-returnable. When the new mattress didn't seem to work too well once I actually tried sleeping nights on it, this was making me reluctant to spend even more money trying another mattress. I reminded myself that the $1400 was a sunk cost rather than a future consequence, and didn't change the importance and scope of future better sleep at stake* (*occurring once per day and a large effect size each day*).	Date of last example: □ Never □ Today/yesterday □ Last week □ Last month □ Last year □ Before the last year

4. What you do when you find your thoughts, or an argument, going in circles or not getting anywhere.

a) I try to find a concrete prediction that the different beliefs, or different people, definitely disagree about, just to make sure the disagreement is real/empirical. (*Recent example from Michael Smith: Someone was worried that rationality training might be "fake", and I asked if they could think of a particular prediction they'd make about the results of running the rationality units, that was different from mine, given that it was "fake".*)	Date of last example: □ Never □ Today/yesterday □ Last week □ Last month □ Last year □ Before the last year
b) I try to come up with an experimental test, whose possible results would either satisfy me (if it's an internal argument) or that my friends can agree on (if it's a group discussion). (*This is how we settled the running argument over what to call the Center for Applied Rationality—Julia went out and tested alternate names on around 120 people.*)	Date of last example: □ Never □ Today/yesterday □ Last week □ Last month □ Last year □ Before the last year
c) If I find my thoughts circling around a particular word, I try to taboo the word, i.e., think without using that word or any of its synonyms or equivalent concepts. (E.g. wondering whether you're "smart enough", whether your partner is "inconsiderate", or if you're "trying to do the right thing".) (*Recent example from Anna: Advised someone to stop spending so much time wondering if they or other people were justified; was told that they were trying to do the right thing; and asked them to taboo the word 'trying' and talk about how their thought-patterns were actually behaving.*)	Date of last example: □ Never □ Today/yesterday □ Last week □ Last month □ Last year □ Before the last year

5. Noticing and flagging behaviors (habits, strategies) for review and revision.

a) I consciously think about information-value when deciding whether to try something new, or investigate something that I'm doubtful about. (*Recent example from Eliezer: Ordering a $20 exercise ball to see if sitting on it would improve my alertness and/or back muscle strain.) (Non-recent example from Eliezer: After several months of procrastination, and due to Anna nagging me about the value of information, finally trying out what happens when I write with a paired partner; and finding that my writing productivity went up by a factor of four, literally, measured in words per day.*)	Date of last example: □ Never □ Today/yesterday □ Last week □ Last month □ Last year □ Before the last year
b) I quantify consequences—how often, how long, how intense. (*Recent example from Anna: When we had Julia take on the task of figuring out the Center's name, I worried that a certain person would be offended by not being in control of the loop, and had to consciously evaluate how improbable this was, how little he'd probably be offended, and how short the offense would probably last, to get my brain to stop worrying.*) (*Plus 3 real cases we've observed in the last year: Someone switching careers is afraid of what a parent will think, and has to consciously evaluate how much emotional pain the parent will experience, for how long before they acclimate, to realize that this shouldn't be a dominant consideration.*)	Date of last example: □ Never □ Today/yesterday □ Last week □ Last month □ Last year □ Before the last year

6. Revising strategies, forming new habits, implementing new behavior patterns

a) I notice when something is negatively reinforcing a behavior I want to repeat. (*Recent example from Anna: I noticed that every time I hit 'Send' on an email, I was visualizing all the ways the recipient might respond poorly or something else might go wrong, negatively reinforcing the behavior of sending emails. I've (a) stopped doing that (b) installed a habit of smiling each time I hit 'Send'* (*which provides my brain a jolt of positive reinforcement*). *This has resulted in strongly reduced procrastination about emails.*)	Date of last example: □ Never □ Today/yesterday □ Last week □ Last month □ Last year □ Before the last year
b) I talk to my friends or deliberately use other social commitment mechanisms on myself. (*Recent example from Anna: Using grapefruit juice to keep up brain glucose, I had some juice left over when work was done. I looked at Michael Smith and jokingly said, "But if I don't drink this now, it will have been wasted!" to prevent the sunk cost fallacy.*) (*Example from Eliezer: When I was having trouble getting to sleep, I (a) talked to Anna about the dumb reasoning my brain was using for staying up later, and (b) set up a system with Luke where I put a + in my daily work log every night I showered by my target time for getting to sleep on schedule, and a — every time I didn't.*)	Date of last example: □ Never □ Today/yesterday □ Last week □ Last month □ Last year □ Before the last year

c) To establish a new habit, I reward my inner pigeon for executing the habit. (*Example from Eliezer: Multiple observers reported a long-term increase in my warmth / niceness several months after... 3 repeats of 4-hour writing sessions during which, in passing, I was rewarded with an M&M (and smiles) each time I complimented someone, i.e., remembered to say out loud a nice thing I thought.*) (*Recent example from Anna: Yesterday I rewarded myself using a smile and happy gesture for noticing that I was doing a string of low-priority tasks without doing the metacognition for putting the top priorities on top. Noticing a mistake is a good habit, which I've been training myself to reward, instead of just feeling bad.*)

Date of last example:

□ Never

□ Today/yesterday

□ Last week

□ Last month

□ Last year

□ Before the last year

d) I try not to treat myself as if I have magic free will; I try to set up influences (habits, situations, etc.) on the way I behave, not just rely on my will to make it so. (*Example from Alicorn: I avoid learning politicians' positions on gun control, because I have strong emotional reactions to the subject which I don't endorse.*) (*Recent example from Anna: I bribed Carl to get me to write in my journal every night.*)

Date of last example:

□ Never

□ Today/yesterday

□ Last week

□ Last month

□ Last year

□ Before the last year

e) I use the outside view on myself. (*Recent example from Anna: I like to call my parents once per week, but hadn't done it in a couple of weeks. My brain said, "I shouldn't call now because I'm busy today." My other brain replied, "Outside view, is this really an unusually busy day and will we actually be less busy tomorrow?"*)

Date of last example:

□ Never

□ Today/yesterday

□ Last week

□ Last month

□ Last year

□ Before the last year