2018-01-03

Superhuman Meta Process

This is a more-or-less complete blueprint of the process I'm currently using to run my life on the meta level. It is organized around four major principles, each of which represents a non-trivial design decision. This means that you could negate each of them, and get something arguably sensible. The minor points serve to explain further, and to map out some consequences of taking the major principles seriously. For me, they are also cached thoughts that help me make decisions.

Many of the minor points, fully expanded, would be enough material for a separate blog post. However expanding them to that level would be quite pointless. If you cannot extrapolate them on your own, you probably shouldn't implement them anyway. In that case, you can probably think of some *other* set of principles that would be a better fit for you. One of the main messages I'd like to communicate here is that *designing your meta process is worth the effort.*

[Added later:] Many people asked me what exactly I mean by "meta" and "meta process" in this context. Let me clarify: I consider it thinking on the ***meta level*** *when I think something like "what trajectory do I expect to have as a result of my whole brain continuing to function as it already tends to do, assuming I do nothing special with the output of the thought process which I'm using right now to simulate myself?". This simulation obviously includes all sorts of everyday changes to my brain, including acquiring new memories, habits, and preferences. However the key question is, am I reflectively consistent? Or: do I endorse the most accurate simulation of myself that I can currently run?*

The ***meta process*** *is what happens when I want to make sure that I always remain reflectively consistent. Then I conjure up a special kind of self-modification which desires to remember to do itself, and to continue to hold on to enough power to always win. I aspire to make this meta process an automatic part of myself, so that my most accurate simulation of any of my future trajectories already automatically includes self-consistency.*

For the sake of brevity, I am writing everything as black-and-white and seen from my personal point of view.

- **Meta process is life.** It's OK to mess up on the object level. However, there is no way to recover from a broken meta level, except by luck. And luck is not enough.
    - **Transparent meta.** This meta process should be public, codified, and teachable. It has nothing to do with keeping the object level confidential.
    - **No second meta.** The meta meta process is just letting the meta process naturally act on itself.
    - **Judge the meta process.** Whether it's other people or organizations, their meta process is what you want to know (regardless of which parts they publicize). Their object-level results are interesting only insofar as they are evidence about their meta process.
    - **Seamless cooperation.** Design your meta process so that it benefits you when you are alone, and also automatically

Blog Archive

January 2018 (3)
December 2017 (3)
November 2017 (1)
October 2017 (2)
September 2017 (1)
August 2017 (2)
July 2017 (1)
May 2017 (2)
April 2017 (1)
March 2017 (2)
January 2017 (2)
November 2016 (1)
October 2016 (1)
September 2016 (2)
August 2016 (1)
April 2016 (1)
March 2016 (1)

More by SquirrelInHell

- AI Safety Comics
- Android Apps
- Be Well Tuned
- Rationality Updates

cooperates when there's an opportunity to do so. Then scheme to be around the people you want to work with.

- **Difficulty maintenance.** Choose your object-level challenges strategically, to maximize how much you learn and grow. However, your meta process must always stay safely within the basin of convergence.

- **Series of expansions.** Everything you do might fail or be abandoned at any time. Strongly prefer plans which are still beneficial if interrupted in the middle.

- **Values flow outward.** You want everything that happens to happen because of your deep values (this might, of course, include other beings achieving their values). Everything else is a bug. Every bug is critical.

  - **Full stack action.** Everything below your top-level values is a tool in their service. Tools are not intrinsically good or bad, but you can make them more or less useful. Useful tools can be stacked together, so that actions propagate far without distortion.

  - **Cooperate not control.** If other people have compatible values, they want to cooperate with you too. If they are adequate, you can't manipulate their values, and you gain nothing by trying.

  - **Reject invest-y power.** Some kinds of power increase your freedom. Some other kinds require an ongoing investment of your time and energy, and explode if you fail to provide it. The second kind binds you, and ultimately forces you to give up your values. The second kind is also easier, and you'll be tempted all the time.

  - **Do only what you want.** Having inconsistent preferences is a bug, and it's in your own best interest to resolve it. Afterwards, there is no need to hold yourself back. If you decide to be nice and share your powers, do it because it's good for you.

  - **Don't ask for permission.** However, when possible, ask for information, and ask to have your models looked at. You'll avoid many mistakes and increase your utility.

  - **Don't apologize. Update.** You aren't in this for emotional comfort. By apologizing to someone else who plays for real, you are insulting them by suggesting that emotional comfort is what they are after.

  - **Automatic respect.** If your models predict you can't get away with doing something, you won't do it. The burden is on others to be adequate at seeing through your motives. There is no need to invoke morality or any special principles here.

  - **Trust only yourself.** The burden is on you to have adequate models. If you think someone will cooperate because it's aligned with their real values, you don't need trust. If you turn out to be wrong, that's your problem. Screening people is always on.

  - **Mental security.** As a human, you don't actually have consistent values, and it takes time to figure out what they converge to. The world is highly optimized to extract energy from you by distorting your thoughts, and making you confused about your values. You cannot trust yourself until you are adequate at seeing through threats in real-time.

- **Aim for full adequacy.** Do not (internally) compromise. Discard inadequacy without second thought. Do not (internally) bow your head to Moloch. Take

ideas to their conclusion.

  - **Everything for your own values.** However, use an adequate decision theory. Neglecting the meta process is two-boxing, and you must not do it even if it locally looks like a thing to do.
  - **Luck is not enough.** It is not enough even if it had already happened. Relying on luck exposes that there is a mission-critical piece which you haven't yet mastered. It'll get you sooner or later.
  - **A plan is not enough.** Do not trust any human (including yourself) to do anything in the future, unless all the options available now have been adequately exploited.
  - **Optimization never stops.** Avoid one-time effort if at all possible. Aim for long-term stability of the process that generates improvements. There is no room for the psychological comfort of certainty.
  - **Infinite inferential reach.** It's not enough to be one or two inferential steps ahead of everyone else. Learn how to build towers of knowledge which can fully support their own weight, so that you can build ever higher. Do not pause as you create whole new disciplines and branches of science.
  - **Every bug is critical.** Each unexpected error, no matter how small, is a canary in the coal mine for some bigger issue. The bigger issue will get you if you don't deal with it right now.
  - **Gain time by taking the time.** You know very well that your time is limited, and you'll constantly be tempted to skip ahead by relaxing your standards. However, each case will be a terrible mistake.
  - **Security from the start.** If your process is not secure now, it won't be secure later when you need it to be. Talk in person. Walk outside. Take notes on paper. Beware of consumer electronic devices.
  - **No gawking at adequacy.** When you start getting some things right, pretty soon most of humanity will look like a bunch of clowns rolling around in the mud. Moving on.
  - **No falling in love.** Being attracted to someone is a sign that your mental security is compromised, and that they are more adequate than you in some respect. Treat it as an important bug report.
  - **One strike and you're out.** If you think you can cheat without getting caught, you should do it! However, if it becomes known that you have sabotaged some important value, there will be no explanation that makes it OK in front of other people who play for real.

- **Update mental software.** The meta process is implemented by gradually self-modifying in the direction of needing less meta-level correction.
  - **Everything is a skill.** You can learn each of them, but it takes time, so be strategic. Stealing skills from people similar to you is faster. If you think something is not a skill, it means you don't understand it yet.
  - **Every skill is a mental skill.** Every skill tree can be traced back to what happens in human minds. Generic mental skills have exceptional return on investment, so make sure to grab them first.

- **Integrate your mind.** Your subconscious processing and emotions are more "you" than your stream of consciousness. You'll not get anywhere if there's any tension on the boundary.
- **Intentions don't matter.** No one cares about the narrative you are telling yourself about your actions. If your mind is not integrated, that's your problem. Take responsibility for all of yourself.
- **Install automatic processes.** The amount of things you can consciously track is very limited. You won't get anywhere unless you can self-modify on the spot and have your brain handle everything in the background. Do it routinely.
- **Refactor your perception.** Fully integrating new understanding feels from the inside like seeing the world differently. Do it routinely.
- **What kills you is blind spots.** It's OK to take big risks knowingly. But if you have a blind spot, you cannot know the upper bound of the risk. You should stop in your tracks at the tiniest suspicion of having one.
- **No dirty imports.** The world will often offer you knowledge in a package deal. However, you must break down every new piece of understanding into its basic components, and own each of them. Gain time by taking the time.

No comments:

Post a Comment

Newer Post Home Older Post

Subscribe to: Post Comments (Atom)