How to make someone who has a shot at creating technical foundations for beneficial superintelligence:

1. **No-holds-barred Gears Thinking:** "Able to see and act on models outside the social. Able to seek gears, and to stay in contact with gears, without needing social 'permissions'."**.**
   Story:
   - As a person aiming to take aim at the whole goddamn AI alignment problem, I had better act as though I am in fact allowed to tackle that problem, because I don't need any added social anxieties or similar ("do I really have a right to try to solve this problem?") impeding my focus on the task. And I had better build my understanding of that problem out of gears and not out of ideologies or self-justifications or similar, because it is gears that hold the keys to the universe's cheatcodes.

   b. **"Experts" as evidence, not as social status barriers.** (repeated below)
      - Looking at evidence, and the causal processes that produce "expert opinion", not at the social credentials.
        - **Story**: As an AI alignment researcher, I need to think clearly about "expert opinion," so that I don't restrict my "right to reason" in cases where I lack the conventional social credentials, and so that I can use the processes of academia/etc. as an aid to forming my own inside-view models.

2. **Drive to Actually Win.** (Virtue of the void / cutting through / creating the mindset of actually aiming to cut the enemy)
   Story: As an AI alignment researcher, I need to *need* to win against AI-related xrisk, so my heart and shower thoughts and mind can be at the disposal of the problem instead of merely {carrying out dutiful motions when hand-scripted by system 2}.
   (Focused grit is one exercise here, but I bet we could make others)
   a. **Finding a worthy enemy.** (e.g. AI risk.)
      - **Story**: As a person wishing to acquire skills such as those on the list, it is extremely helpful to have viscerally noticed a worthy enemy (aka "Something to Protect"), so as to have a felt need for the skills.

   b. **Deep caring without forcing.** Skills for turning yourself into the sort of person who can have deep caring without ever incurring internal resentment or burnout-risk, or losing internal hit-points in that way. (Elaboration: A bit like Kenzi's "do what s1 wants and what s2 wants; give both parts *all* the controls" thing. Knowing you're allowed to care. Nate's business about re-anchoring at odd scales, sense of the absurd, …)
      - **Story**: Human S1s are good at grasping relationship goals, career goals, and other "human-scale" goals that feed well into our EEA drives. AI risk is not such a goal. Humans who attempt to just paste in a goal that is

alien to their system 1's tend to work poorly or burn out. And so, as a human aiming seriously tackle this problem, I'll need the ability to do something else -- to access "deep caring without self-sacrifice"; to help my being turn toward a deeper goal that it yearns to care about but that is yet alien to it, while still not forcing things. I need to do this so that I can bring real effort to bear in a sustainable, and "grows over time" way.
        - (Anna has a lot of detailed thoughts here if you want them. See also all of Nate Soares's blog posts.)
    c. **Deep pain/restlessness/need to "do".** (Perhaps all humans care on reflection about e.g. large near-term human extinction risks, or their own ambition for that matter, but have it blocked off somehow; in any case the idea here is to activate this.)
        - **Story**:
            - As an AI alignment researcher, I must access deep energy, or a deep well of impetus to action, so I can head into a near-impossible, decades-long task with actual velocity, and can maintain that velocity for decades, instead of simply making motions at things.

3. **Whole-problem Engagement:** Ability to not just bounce off the big "actually solve AI" problem, and to instead stay engaged and wrap your mind around the problem as a problem.
   Story: As a person aiming to tackle the whole of AI alignment, I need the ability to correctly see myself as someone who has some shot at this, so I can in fact "shut up and tackle the impossible". This ability is made up of a combination of skills (so I can see I have a shot at it) and insane powers of anti-flail.
    a. **Ability to actually try on problems an untrained human would run away screaming from.**
        - **(7) Anti-flail**
            - **Story**: As an AI alignment researcher, I need to be able not to flail, because otherwise this is totally the sort of problem that a person would run away screaming from. And I need to not do that!
        - **Intellectual courage and robustness, including in the face of weird.**
          Story: The issues involved in AI risk can be profoundly disorienting. I need to reliably: keep thinking anyhow; remain a good person; trust my models as much as they should be trusted and no further; and just basically display "hobbit virtue" when dealing with things far removed from the human scale. I need this so I can keep my ability to think and to act well as I tackle the AI alignment problem.
            - **Moral and ethical stability.**
                i. Story: As a person who will be exposed to extremely high-stakes, extremely confusing

questions that ought in some sense to belong to all of humanity (if successful), I desperately need moral and ethical stability so I can reliably act well anyhow. (cf Hobbit virtue.)

- **Epistemic stability.**
  i. E.g., don't literally go insane in the face of simulation arguments and similar.
    - Story: As a person who will wrestle with many of the biggest gaps in our current understanding of how the world works, I need the ability to watch my ontology change or teeter underfoot, and to stay sane anyhow.
- **Epistemic courage**: Willingness to go against the conventional wisdom if a good argument suggests doing so. Willingness to believe that hard problems are solvable, and to dedicate a lot of time to attempting to solve them.
- **Comfort with "weirdness"** -- whether that's social weirdness, or epistemic weirdness. Not dismissing a line of inquiry solely because of weirdness. And understanding how to update your credence based on degree and type of weirdness.
  i. **Story**: As an AI alignment researcher, I need to be comfortable with weirdness, so that I don't reject important (but not yet mainstream, or counter-intuitive) ideas.

- **(repeat) Looking at evidence, and the causal processes that produce "expert opinion", not at the social credentials**
  - **Story**: As an AI alignment researcher, I need to think clearly about "expert opinion," so that I don't restrict my "right to reason" in cases where I lack the conventional social credentials, and so that I can use the processes of academia/etc. as an aid to forming my own inside-view models..
- **"No Halo of destiny"** / you don't need a Hogwarts letter
  - As an AI alignment researcher, I need to viscerally grasp that I [making progress is a matter of what I learn how to get physics to let me do, and not a matter of social permissions], so that I can tackle problems I haven't earned the credentials to tackle.

b. **Research relevance-detection: Ability to sift out what is/isn't central.**
Story:

- **Good taste**: Aspects of this include sensitivity to simplicity, beauty and elegance, and (perhaps the same thing???) to what kinds of arguments and models are likely to lead in fruitful directions and what others are

likely to be dead ends. I mean by "taste" more or less what Paul Graham does.

- **Story**: As a student of AI risk, I need to have good taste so that I can spend my time on lines of inquiry that are likely to be fruitful.

- [domain-experience? what else?]

c. **Research-strategy. Drive and skills for creating a systematic sense of the total problem-map, and of which pieces to tackle when.** Skills for moving forward / noticing forced moves / etc. Skills for plotting a path and pursuing it. Strategicness about intellectual research. Ability to break things into pieces, track the whole gameboard, make progress, etc. Goes hand in hand with 'b' to allow you to hold a map in your head in a useful fashion.

Story: As an AI alignment researcher, I need to think strategically about the whole research landscape so that I can work productively on the highest leverage parts of the problem.

- **Drive to systematize.** To form a map of what does/doesn't need tackling, of what may be missing from your considerations, etc.
  - **Story**:As an AI alignment researcher, I need to have a drive to systematize so that I can turn a big fuzzy problem into concrete, tractable avenues of attack.
- **Creating terminology that helps with thinking** - pithy abstractions that are accurate enough to mostly be relied on, or that clearly highlight their limits.
  - **Story**: As an AI alignment researcher, I need terminology that helps with thinking, so that I can simplify complex topics in ways that allow for productive thinking and don't leave hidden traps.
- **Jumping beyond yourself**: Be aware of when your thinking on a topic is no longer progressing. Figure out why & how to jump outside the pattern.
  - **Story**: As an AI alignment researcher, I need to notice when my thinking on a topic is no longer progressing, so that I can figure out what pattern I've been repeating and how to jump outside it.
- **Disjunctive reasoning:** When you know that "A or B" is true, fully inhabit each possibility separately ("What would a world with A look like? What would follow if A is true? … And what a world with B look like?)
  - **Story**: As an AI alignment researcher, I need to be able to fully inhabit each possible scenario (separately) when there are multiple possibilities, so that I can avoid the failure mode "We don't know the answer to A yet, therefore I can't think about X at all"
- **Understanding dominated strategies and "forced moves"**
  - **Story**: As an AI alignment researcher, I need to understand forced moves because we are unlikely to resolve uncertainty in most of the landscape of the AI alignment problem, and we need to be able to identify good actions despite remaining uncertainty.

- **Explicitly label blank spots on your map**. Have some idea of what goes in and out of them, but expect your notion of the API to change too once you understand what's in there. Give them silly names like 'magical reality fluid' so you don't start believing in them the way some people believe in 'modal realism'.
- **Get to terms with partial progress.** Learn to enjoy shouting "Progress" instead of "Victory" because you may not be shouting victory for 20 years.
- **Maintain a concrete examples library:** Have enough central examples (hypothetical scenarios about how things might play out) about the big problem that you can check your next story against the examples, even if you don't have very much of a *theory* about the whole big problem, so long as you have example scenarios.
- **Have an actual big picture.** Put it up on a whiteboard sometime. Draw an awful Mentifex-style diagram of what you think the boxes are inside an AI and what the boxes are doing. Make a Workflowy outline containing a list of what you think are all the top-level unsolved problems in AI. It's going to look really awful. You're allowed to edit it after that. This is not your promised research plan that you have to stick to for the next 3 years, it can be revised anytime you like and it *will* be wrong, you're just doing it so that you have a big picture at all.
- **Constantly assess what you do and don't think you understand.** Do you understand anthropics? How well? Do you think you understand the problem of unforeseen maximums? How well?
- **Constantly ask 'what if we actually solved this'** - one way of noticing that you forgot to have a big picture model, is when you can't answer the question of how the big picture model makes a function call to the thing you're using now. Be comfortable with working on something that seems very far from finishing the big problem, but be uncomfortable with something where the story about how it ends up being relevant involves "And then a miracle occurs" or "And then this problem is similar to X". Have a roadmap and use that roadmap to try to predict which work now will end up being relevant in 20 years.
- **Be okay with uncertainty and holding two different possibilities in your head at once.** make sure you're only putting in that effort if you feel real uncertainty (plausibility in both cases) and not just a sense of obligation or not being allowed to come to a decision.

d. Intellectual maker skill / creativity / etc.

- **Original seeing.**
- **Steel manning / rationalizing up new important skills near the old ones.** (cf.: Generalizations of prehindsight.)
    - **Story**: As an AI alignment researcher, I need to be able to steelman other people's views, so I can recognize good pieces of thinking even when they're bound up with wrongness, since none of us is likely to get all of it right on our own.

    - Anna thinks *most* of her power comes from strategic use of steeling.
- **Intellectual courage.**
    - Ability (and willingness) to express vague, not-yet-crystallized impressions.
        i. **Story**: As an AI alignment researcher, I need to be able to express vague impressions, because the field is early-stage and the ideas that will eventually *become formal pieces of the field are still inchoate.*
- *More skills to be elaborated:*
    - *Writing first drafts with them being shielded from view, so that you can access all your thoughts*
    - *Having big models and big dreams*
    - *Artistic taste*
    - *Sense of what will become relevant later*
    - *Pressurepointing. (See below.)*
    - *Being a Fox made out of Hedgehogs: (Eliezer: The whole thingy I wrote on Facebook about the bad thing that happens if you think being a Hedgehog is bad; have big theories that you care about, be able to modify them or toss them out the window, make sure you don't have just one big theory: https://www.facebook.com/groups/674486385982694/permalink/784664101631588/)*
    - *Strategic use of steeling / rationalization. "If there's an amazing thing you can do with honey, cocoa powder and peanut butter, how would it go?" (aka: Generalizations of prehindsight).*
    - *You got here by asking a Hamming Question, right?*
        i. *finding the right thing to work on*
            - *scope sensitivity*
            - *marginal impacts*
            - *being able to perceive inadequacies even when everyone's agreed not to work on them*
    - *Naming things / systematicity*
    - *Never even start to do a major undertaking that is boring (esp. if a research/writing/creative undertaking)*
        i. *but if you do, notice that you're bored*
        ii. *if you have to work on it anyway because it's super important, how did it end up being boring? something must be wrong. at least if we're dealing with conceptual stuff like math. labwork can be boring so long as the larger problem isn't boring.*
    - *Grand Organization:*
        i. *systematicity so you want to name all the things*
        ii. *elegance so you don't want too many as top-level categories (what stops you from inventing 7 Tegmark Levels instead of 4)*
        iii. *Relevance for moving things in and out of categories in your outline ("that doesn't fit here... doesn't fit there either... I guess it goes to the top level, now can I figure out a supercategory or does it just stay there?")*

        - *Keeping a running list of everything you want in case you might be able to get it someday (as applied to theories and their theoretical qualities)*
            i. *A sense of urgency about things that are currently broken*
            ii. *A list of things that are currently broken*
            iii. *Ability to mark things as broken even when people are not currently working on them or paying attention (e.g. journal paywalls before journal paywalls were a big issue).*
- **Skills for internal dialogue/having beliefs**
    - **Identifying your "true rejection"** (by doing a quick thought experiment before you put a justification forward: "if I found out [my justification] was false, would I still believe X?")
        - Story: As an AI alignment researcher, I need to be able to identify my "true rejection" so that I can hone in on the small set of considerations that are actually core relevant to the issue, and that might actually change my mind, instead of constructing and talking about a mass of shadows and distractions.
    - **Internal "2-player 20 questions".** Skills for figuring out the cause when part of you disagrees with another part of you.
    - **Noticing confusion.** Using TAPs to refine one's ability to notice confusion.
        - Story: As an AI alignment researcher, I need to notice when I'm "hiding the 'could' under the rug", vs when I'm making progress, so I can avoid heading off into random pointless fake-thought.
    - **Use the native architecture:** Correctly harnessing Curiosity, Taste/Aesthetics, Playfulness, Strategicness, and other pieces of our hardware:
        - Getting back into "curious / actually trying to figure it out" mode
            i. Story: As an AI alignment researcher, I need to be able to get curious, and to prevent false explanations and motivated cognition, so that I can keep my thinking in a mode where it makes me smarter and not dumber.
    - Eliezer: **To carry out useful dialogue inside your head you have to unlearn all the terrible, terrible habits you have learned from conducting debates with other people.**
        - It is your responsibility to carry on all of however many parts there are inside the conversation; nobody else is going to take the 'other side'(s) if you don't.
        - You're actually going to need to learn to transcend the whole adversarial paradigm of dialogue, though taking all parts simultaneously is the traditional place to start. a

defender plus a prosecutor does not a Bayesian agent make.
        - you're responsible for the questions of when to criticize, when to continue searching for a critique, when to stop searching for a critique, when to launch a search for a counter-critique
        - ...actually I'm having trouble figuring out what to say about this that isn't just "use Weakpointing on your own ideas".

- **Creating New Applied Math**
    - express vague, not-yet-crystallized impressions
    - relate high-level goals to math properties
    - relevance: how would the plot change if we changed this equation?
    - notice when progress is not occurring: flag, analyze introspectively for 30 minutes including typing or conversation, regroup if necessary
    - notice confusion
    - notice when you've just seen something very important and ought to stare at it a while
        - Anna: no universal compression because pigeonhole
        - Eliezer: minimum message length
    - Feynman's cup, 'look at the water'
    - Lakatos's "Proofs and Refutations", Polya "How To Solve It"
        - Polya: teaches to go back and look for a more elegant proof after proving
        - Lakatos: teaches how intuition comes from examples, how proofs often contain hidden assumptions, and how proofs often need to be rescued by stating the assumptions. see also "Modularity".
    - Rationality-as-winning, illustrated in Newcomb's case
        - Being able to question the criteria
        - Being able to invent the criteria
    - **always try to visualize what the equation does**
        - e.g. by substituting 4 for all appearances of x, and maybe even 4 apples for all the 4s

e. Basic knowledge of the AI risk problem. Starting skills for accessing the research.

4. **PressurePointing.** (finding the pressure point where a theory will give; system-1 noticing what is most wrong with a thing). (This is overwhelmingly important in practice

to value alignment work, because in fact almost nothing solves the whole problem for advanced agents, and almost everyone else is going around claiming that their latest trick does so.)

a. if you're going to rationalize something, it helps your mental health a whole lot to be rationalizing a true thing instead of a false thing. in value alignment theory, "this doesn't solve the whole problem" is always true and "this solves the problem" is always false, but that doesn't mean you can't be wrong about how or more importantly where a solution fails, which means critiques don't automatically win or get right-of-way.
b. so you have an idea. you're trying to find the weak point in it, a place where it fails, or the first place where it fails, or the worst way that it fails, or even (for explanatory purposes) the most understandable way that it fails.
c. maybe your brain just generates something. if so, go to the 'is that critique actually true?' step.
d. has your brain not generated a critique yet? go to one of your central examples, see how the idea plays out in it. ask how this idea would work for building a paperclip maximizer or diamond maximizer, or whatever the appropriate central example is here.
e. if you see what seems like a concrete failure, generalize it abstractly and then go to the "Is that true?" step
f. suppose somebody told you as a solid fact that the thingy had failed. what would you update to believing in that case?
g. if there's anyone who wouldn't believe the thingy, what would they say? try both the 'steelperson' and 'Ideological Turing Test' version of these; these are different skills.
h. okay, you have a critique. is it actually true?
i. try to generate counter-critique, imagine being told that the critique had failed, what would somebody who believed the original thingy say, etcetera
j. Check Sense of Fit. check common sense. check common examples.
k. is there a way to repair the original idea that withstands the critique? what would the obvious repair be? (also consider this both in steelperson form and in ITT)
l. does the 'repaired' idea still have the critical properties of the original, or were those properties lost? can you repair the original idea in a way that preserves the properties?
m. was your critique the most important flaw, the one that would be hardest to repair, or are you seizing on a technicality? (need for closure here, or the desire to have 'found a critique' or to be done with the critiquing step, are your enemies.)
n. strong phenomenal resemblance here to Polya's "Can you improve the proof? Can you see it a glance?" step of mathematical reasoning.
o. the fundamental trick of Weakpointing is to have it not be a foregone conclusion that everything fails, nor that every debate continues forever. at some point you can't find any more counter-critiques, you can't repair the original, you stare at the thing and realize that this is just true and you shouldn't expect to find a strong

critique of it. for the skill to be useful, it has to be discriminative between solid and unsolid ideas, without (a) all things seeming false because you are too good at critiquing, (b) all things seeming true because you are too good at steeling and rationalization, or (c) nothing ever finishing because you can always come up with one more counter-counter.

5. **Actual rigor.** (Aka: defense against false “knowledge”).
   Story: As an AI alignment researcher, I need rigor so that I can root out mistakes and confusions in my thinking and build more and more refined models.
   a. **Noticing the implications of your beliefs**
      - **Story**: As an AI alignment researcher, I need to be able to notice implications of my beliefs, because AI alignment reasoning requires following multi-step chains of logic.
   b. **Modularity.** When you disagree with someone, be able to hypothesize a premise on which you might disagree. Then ask them if you disagree on that.
      - When somebody else says "I think A and B, and A and B imply C. Do you disagree with C, and if so do you disagree with A, B, or the implication?" be able to answer them, god damn it. be Robin Hanson, not Tyler Cowen.
        - classic test: Bostrom's Simulation Argument (as distinct from the Simulation Hypothesis, thank you very much). Hanson used cryonics on Cowen. something more 'normal' here than SA would be nice.
        - (this is a test of whether you can use somebody else's modularization. coming up with your own modularization is harder.)
      - Know what to do when faced with a False Dichotomy in somebody else's bad modularization instead of panicking; like suggesting different categories, omitted items, third alternatives, etcetera. e.g., figure out what assumption underlies the appearance of Dichotomy, then argue with this assumption
      - possibly: ask what prior facts would have to be different about the world for the Thingy to end up false, then that property being not-different is an assumption for the Thingy to be true.
        - what is the minimum departure from reality that invalidates your idea?
        - "I think blah is true! For blah to be false, you might suppose blah-blah." The more somebody might actually be disagreeing with you because they believe blah-blah, the better you're doing at blah-blah selection.
        - on Monday, we live in a universe where the best explanation for global warming is humans. on Tuesday, we live in a universe where it isn't. what's the most probable or most reasonable thing

that could've been true in the background on Tuesday to make that be the thing that ended up being true?

c. **solution checking**
    - does it solve the entire problem? what are the background assumptions required for this to be solving the entire problem? state explicitly what they are.
    - does your explicitly labeled "????" step actually contain the whole problem over again? (e.g., chess player that has a 'good move detector' and a bunch of other boxes like 'memory storage' that are not really the hard part of the problem)

d. **Seeking empirical tests.** Designing cheap experiments. Googling. Popperian virtue.
    - **Story**: As an AI alignment researcher, I need to seek empirical tests in order to take advantage of opportunities to better entangle my beliefs with reality and resolve uncertainty quickly.

e. Modeling skills. Ability to form simple mathematical models, and to mistrust them.
    - Familiarity with basic math and with basic math models of some common domains (e.g., evolutionary biology; microeconomics)
    - Model combination. e.g. in http://lesswrong.com/lw/hzu/model_combination_and_adjustment/
        - **Story**: As an AI alignment researcher, I need to be able to think clearly about different models, since different ways of thinking about the problem lead to different results.

f. **Ideological Turing tests.** Forcing yourself not to dismiss a person's point of view until you can pass their ideological Turing test to that person's satisfaction.
    - **Story**: As an AI alignment researcher, I need to practice the discipline of forcing myself to pass the ideological Turing tests of the best folks who disagree with me, so as to have at least a partial saving throw against being stuck in my own blindspots.

g. **2-player 20-questions**, for honing in on the causes of a disagreement.
    - **Story**: As an AI alignment researcher, I need to be able to rapidly hone in on the causes of disagreement with other researchers so that we can update from each other.

h. **Staying oriented on the real question**. Getting back into "curious / actually trying to figure it out" mode.
    - **Story**: As an AI alignment researcher, I need to be able to get curious, and to prevent false explanations and motivated cognition, so that I can keep my thinking in a mode where it makes me smarter and not dumber.

i. **Understanding understanding.** Capacity to clearly distinguish between things you understand and don't understand. Not stopping at the level of "someone gave me a proof and I can check the steps," but ideally at the level of "I have a

detailed enough mental model of how this works that I could have independently generated this claim."

- **Story**: As an AI alignment researcher, I need my map of the field to be full of moving pieces that I can play around with, not just a static collection of facts, so that I can create new research and correct mistakes.
- **Noticing things that masquerade as reasoning** (patterns of thinking that aren't Bayesian or strategic, but which feel like reasoning)
  - **Story**: As an AI alignment researcher, I need to notice things that masquerade as reasoning, so that I can avoid allowing such things to distract me or to distort my mental landscape
- **Righting a "wrong question"**
  - A. Noticing the limits on what is well-explained; distinguishing deep understanding from speculative explanations & isolated facts
  - **Story**: As an AI alignment researcher, I need to be able to notice and dissolve "wrong questions", so I can avoid both {throwing useless effort at such questions}, and {training myself to ignore the feeling that I'm still confused}.
- **Noticing discordant evidence**: being able to go from "X is clearly right because it seems true to me, done" to "But Y also seems true, at least to some people. I'm not done until I can explain not just why X is true, but why Y is false."
  - **Story**: As an AI alignment researcher, I need to notice discordant evidence so that my thinking does not get stuck in a model which seems to explain things better than it does.
- …?
  - [noticing things that feel painted-in / made-up, vs things that feel forced. Gears thinking.

j. Having beliefs.

- **Accessing the causes of an intuition:** Using thought experiments to get at the causes of an intuition, and persisting until you have a causal model of the process generating the intuition. ("I feel as though the car is dangerous to drive; if it wasn't making this noise, would I still feel that way?")
  - **Story:** As an AI alignment researcher, I need to be able to use thought experiments to understand the causes of my intuitions, so that I can dialog effectively between my intuitions and my verbal/explicit reasoning (since each may see distinct and useful parts of the problem).
- **System 1 "gut checks":** Doing "gut checks" of your professions using System 1. For example, "Do I really believe X?" does it have the feeling of an empty claim or rationalization? Does it "feel" too neat or pretty? How surprised would I be to learn X was false?

        - **Story**: As an AI alignment researcher, I need to be able to do System 1 gut checks, because AI alignment is an abstract topic and it's easy for people to not notice when their professions don't make sense or aren't supported.
    - **Dialoging between anticipations/inner sim and explicit/verbal models**. Making a habit of player "internal 2-player 20 questions", or finding a quick empirical test or similar, whenever two pieces of you and you disagree. Forming coherent models.
        - **Story**: As an AI alignment researcher, I need to be able to dialogue between my anticipations and my verbal models because they both contain insights that the other is prone to miss, and it is easier to make progress on a problem if they are working on it together.

k. **Making your thinking transparent.** Taking a piece of reasoning you just did, and making a mechanistic model of the descriptive process that seems to have in fact generated it in your mind (tested via thought experiments, etc.), and also (separately) of the normative reasoning that would've yielded an accurate answer in this case. (Special case: building Bayesian habits)
    - **Story**: As an AI alignment researcher, I need to model my own reasoning, because it's much easier to notice bad (and good) features of my reasoning when it's legible. (Also makes it easier to communicate your reasons with other researchers, and thereby Aumann.) (Also is good practice toward understanding and designing an AI, and understanding and designing good rationality skills.)
    - **Building bayesian habits.** (Knowing Bayes' theorem; being able to spot it in action in your own automatic thinking in real life, and to notice when/how your thinking goes astray from the math of the evidence.)
        - **Story:** As an AI alignment researcher, I need to be skilled at modeling both the causes of my beliefs, and the extent to which those causes can/cannot be expected to form accurate models. I therefore need also to know basic templates for modeling accurate belief-forming processes, to facilitate this process. For example, I need to know Bayes' theorem, and to be adept at spotting the unconscious mental motions that correspond to intuitively represented probabilities, likelihood ratios, prior odds estimates, etc., and to be adept at noticing when my brain is/isn't following processes that can be expected to yield accurate maps in simple Bayesian updating situations.

l. **Toolkit-building:** Developing an explicit toolkit of identifiable thinking skills
    - **Story**: As an AI alignment researcher, I need an explicit toolkit of identifiable thinking skills, so that I can think strategically about what skills to use on what problems.

6. **Recursively creating new thinking skills when needed.**
   Story: As an AI alignment researcher, I need to be able to create new thinking skills because making progress on a problem as difficult as AI alignment is likely to require skills that haven't been identified yet, and skills that are hard to learn via straightforward transfer from others.
   a. All of "Actual rigor". Plus:
      - **Story**: As an AI alignment researcher, I need rigor when modifying my thinking patterns so that I don't screw up my thinking.
      - **Alternate Story**: As an AI alignment researcher, I need rigor in thinking about the problem and my own thinking skills so that I can recognize when my current skillset is not up to the task before me and seek out new skills to fill the gap.
   b. **No tools? Make tools!**
   c. **Tutoring skill (/writing skill)**. Aspects include the ability to model other peoples' confusions in enough detail to be able to resolve them with explanations, and the ability to clearly communicate new insights in a way that causes other people to have them too. **Includes "pithification".**
      - **Story**: As an AI alignment researcher, I need the skill of modeling how other people are thinking about relevant topics so that I can develop better models of my own thinking on the topics, and of the topics themselves.
      - **Alternate story**: As an AI alignment researcher, I need skill at explaining things so that I can create collaborators to help me make progress on research. [Note: this may fit better under "6. Effective Collaboration"]
   d. Awareness of how your thinking works on a trigger-action pattern level, and ability to alter it
      - **Story**: As an AI alignment researcher, I need be aware of how my thinking works on a trigger-action pattern level, so that I can improve my default/automatic thinking patterns (e.g., by noticing where they should/shouldn't be expected to produce an accurate map).
   e. Acquiring new thinking skills from other people
      - **Story**: As an AI alignment researcher, I need to be able to dissect and copy any useful thinking skills I see in others, partly so I can have those skills and partly because it'll give me general xp-gain toward the project of being the creator of my own mind, and of whatever new skills the AI problem may demand.

f. **Seek the source:** Taking a piece of knowledge, and asking how you could've invented it. Using this to create new pieces of rationality-art. (http://lesswrong.com/lw/la/truly_part_of_you/)
    - **Story**: As an AI alignment researcher, I need to be able to reverse engineer knowledge, so that I can learn the skill of creating knowledge, and thereby advance the art of AI-relevant rationality.
g. **"Update the zeroth time":** Taking an error, and asking what general-purpose rationality skill or piece of rationality art, if you had had it, would've let you prevent that error. (cf: Val's "Update the zeroth time"; http://lesswrong.com/lw/k0/singlethink/)
    - **Story**: As an AI alignment researcher, I need to reverse engineer errors so that I can prevent myself from making similar errors in the future, especially given that some errors won't be apparent until it's too late.

7. **Effective collaboration**
   Story: As an AI alignment researcher I need to collaborate effectively, because a group of highly skilled people can make more progress on a difficult problem if they are working together, but it can be difficult for people with skills 1-5 to coordinate.
    a. **"Knowing one's place"**: Despite having property #1 ("not knowing one's place"), has also the ability to model the role others in fact regard them as having within the company/team/world's "family system", and to not create ripples there unless on purpose.
        - **Story**: As an AI researcher, I need to understand other people's maps of how my role fits in with others' roles, so that we can take advantage of the benefits of division of labor & relatively predictable roles without interfering with my ability to see the whole strategic landscape and focus on solving the actual problem.
    b. **….** lots of other things.


Calvin and Hobbes
by Watterson
HOOLS
UH OH, I'LL BET HOBBES IS WAITING TO SPRING ON ME AS SOON AS I OPEN THE FRONT DOOR!
I KNOW! I'LL SNEAK AROUND BACK AND SURPRISE HIM!
HEH HEH! THERE HE IS, ALL READY TO POUNCE! WHAT A SUCKER!
I'M HOME!
I'VE GOT TO START LISTENING TO THOSE QUIET, NAGGING DOUBTS.