2017-12-30

Notes on Mental Security

Note: these are some rough, preliminary notes on mental security, defined as the art and discipline of keeping one's mental software free from hostile external influence. Or in other words, having a mind able to achieve its own goals, rather than goals of, say, Moloch or other people which happened to be around. This is supposed to serve as a reference point for discussion and further thought. It's definitely not an explanation or a tutorial.

- Why?
  - this is dangerous business, and not one that human minds are designed for
  - however, necessary to do anything which is far outside of the Overton window of agency
    - if you are doing such a thing, you might just as well ignore the risk, because if you fail at security your cause is doomed anyway
  - of course, security is pointless unless you are convinced that you actually have some integrity/values/thoughts that are worth protecting
  - and it should go without saying that working on security with insufficient self-knowledge is just shooting yourself in the foot really hard
- Detection / Self-Diagnosis
  - Male-Type Threats
    - which people or sources...
      - are you impressed by?
      - do you turn to when in doubt?
      - do you listen to/read very eagerly?
      - always have convincing arguments?
      - seem to always be right?
      - you can't imagine disagreeing with?
    - monitoring beliefs
      - coherent meta-epistemics
      - making beliefs pay rent
      - tracking social reality separately
      - disconnecting beliefs from identity
      - noticing burned beliefs
        - beliefs which are maxed out the scale of intuitive confidence, indicating that the whole scale is miscalibrated
      - wielding artifacts of power
  - Female-Type Threats

        - which people or social strategies...
            - do you tend to find attractive?
            - make you feel safe?
            - make it hard for you to think clearly?
            - are you jealous of?
            - you can't imagine living without?
            - cause you to ruminate a lot?
            - you can't imagine hurting?
        - monitoring emotions
            - what emotional reactions do you have that seem inconsistent with your values?
                - again, should be obvious but: are you by any chance wrong about your values?
            - do they predictably change in certain situations, or around certain people?
            - are you sometimes surprised by your own emotions or actions?
            - do you experience akrasia?
            - do you tend to avoid situations which would make you experience strong emotions?
            - packet capture
                - spend a few days in isolation, watch compromised mental processes try to restore communication

- Prevention
    - learning to see through threats in real-time
        - building models in increasingly difficult situations
        - watching from afar how other people are being pwned
        - replicating offensive tactics to learn about them
        - pentest by a trusted and skilled third party
        - strong meta-rationality to estimate threat levels
    - offense is (some) defense
        - however, this is leaky and fragile
        - might give a false sense of security until you are suddenly out of your league
        - it is common to subconsciously gravitate to this local optimum
        - there are ethical concerns
    - firewall
        - gather evidence about power levels of various treats
        - avoid exposure to known threats above your level
        - keep physical distance if possible

Warning: all the articles linked above potentially lead to lethal memetic infection loads. There isn't much I can do about it.

No comments:

Post a Comment

Newer Post Home Older Post

Subscribe to: Post Comments (Atom)