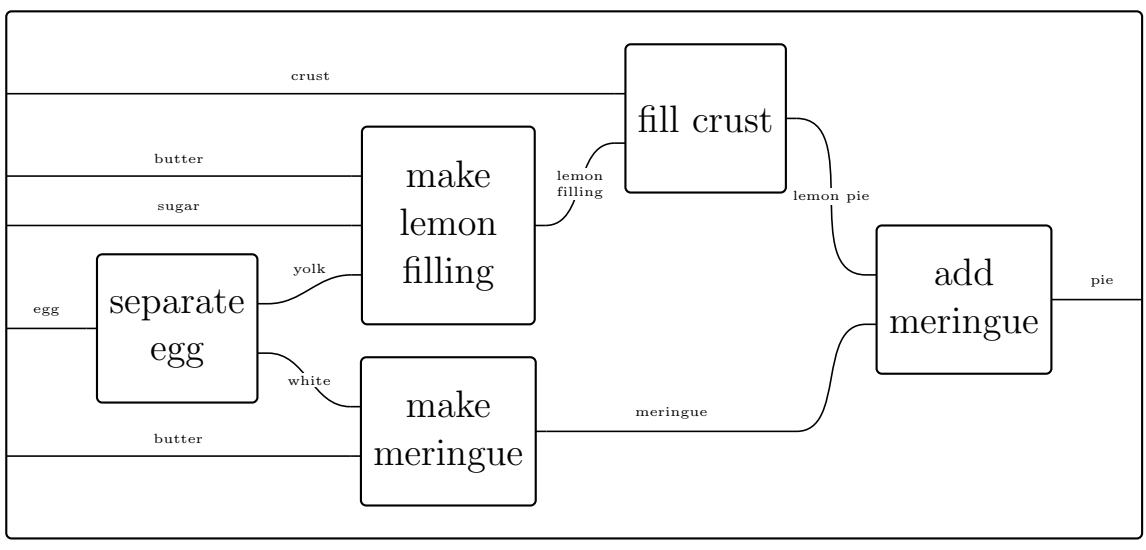
How should one govern a plausible fiction platform?

David Spivak

2024 / 10 / 09

- Plan
  - Describe plausible fiction (5 mins)
  - The planning-proof correspondence (3 mins)
  - A platform proposal (7 mins)
  - Making it real (5 mins)
  - Discussion: (40 mins)
- Plausible fiction
  - Rules
    - Start now, minimal distortion
    - End in good/ok/open future
    - Plausible the whole way (physics, social dynamics)
    - Memetically fit for collaborators
  - Collaboration via gap filling
  - Lightning bolt and stepping stone analogy
  - Draw wiring diagram



- The planning-proof correspondence
  - A math version
    - Mathematical conjectures as "good futures"
    - Lemma factorization as plausible fiction
  - The planning-proof correspondence
    - (Conjecture : Lemma Factorization)]
      ::
      (Outcome : Plan sketch)
    - Prove I can cook dinner.
      - It suffices to prove I have a recipe and the ingredients on it
      - I have a recipe. Prove I have each ingredient.
      - Oops, I don't have carrots; prove I can get to the store and that the store has carrots
- A platform proposal (joint with Quinn Dougherty)
  - In Lean, Coq, Agda, etc.
  - Two sides
    - Public side: contracts = (Username, $bounty_amount, statement)
      - Example: (Clay Institute, $1m, RH), where RH is a Lean term for Riemann Hypothesis
    - Private side: IP = (Username, premise list $[P]$, conclusion Q, proof term $\alpha : P_1 \wedge \cdots \wedge P_n \Rightarrow Q$)
      - Trigger: Once all premises are proven, it fires.
      - If there is a bounty for Q it is paid (split betw. winners)
      - If premise list is empty, you've proved Q; instantly fires.
      - All proven theorems move to public repo as soon as they fire/are paid.
    - Alice has a factorization of RH
      - She writes (Alice, $[P_1, P_2]$, RH, α) privately
      - She writes (Alice, P_1, $100k) and (Alice, P_2, $150k) publicly.
      - Risk: if someone proves P_1 but not P_2, she's out $100k.
  - Why privatize:
    - Cautionary story if proofs were instead public
      - Bob replaces P_1 with equivalent P_1'
      - Now alice is on the the hook for P_2 but both split the RH pot.
    - It's hard to check equivalences, so the machine can't detect this.
- Making it real
  - Task rabbits and AI
    - Bounty for building a deck for my house
      - Contractors / workers factor the problem
      - LLMs might get involved in either math or taskrabbit example
      - Robots?
  - Governance issues: open discussion
    - Will this platform work / be effective?
    - Will it be good for the world? (Misuse, future of human work)
    - Is something like this inevitable?
    - How should you govern it?