At yesterday's AI Safety Unconference presentation on learning biases, Daniel asked me (Remmelt) for examples of focussing on processes (vs. on structures).

I pasted tentative cases below where the process distinction seemed relevant:
–> **Examples for decision-making**
–> **Examples for modelling the outside environment**

Distinction between focussing on (interpreting and forecasting) structures vs. processes:

- *Structures* – We tend to perceive the world as consisting of parts:
  - we represent + recognise an observation to be a fixed cluster
    eg. an observation's recurrence, a place, a body, a stable identity
  - we predict + update on the possible location of this structure
    eg. how likely it ends up present within some linear sequence, geographic surface, physical boundaries, or levels of feature abstraction

- *Processes* – Conversely, some traditional cultures foster a process-based view.
  - they represent + recognise an observed change to be a trajectory in presence
    eg. a transition, movement, interaction, relation
  - they predict + update on whenever, wherever, and so forth, this process may initiate again

(More succinct distinction I came up with later:
- Structure thinking is 'there is this particular locatable thing I can recognise – with particular abilities to move and transform'
- Processes thinking is 'there is this particular repeating dynamic I can recognise, which has a tendency towards merging or separating out stuff' **)**

My explanation of this distinction and why it's important has been counterintuitive to rationalists/EAs.
Do tell me how you would explain it more clearly from your or others' perspectives!

Focus in decision-making:
(Motivated Locomotion vs. Value Assessments Between Points)

1. **Approaches to reinforcement learning**:
   Policy-based RL vs Q-learning
   - choosing best action-reward transitions (?) *vs.*
   - assessing most valuable states to end up in(?).

2. **What in playing games you find rewarding:**
   Infinite vs. Finite Games
   - playing a game of catch: the actions are rewarding in of itself (you associate the reward with the act of throwing or catching, the social interactions between you and your dad, etc.) *vs.*
   - a soccer game: you associate the reward with particular end states (that the ball will be located inside the goal net; the *way* you act towards that outcome doesn't matter as much)

3. **Moral judgements:**
   Deontology vs. Utilitarianism
   - *Deontology* is about willing a *relational process* into or out of existence (e.g. the act of lying ends motivation for lying, since interlocutors no longer rely on each other's honesty), premised upon the value of a generalised social relation.
   - *Utilitarianism* is about assessing between the intrinsic value of *structured identities* (particularly, valances or met preferences) predicted to exist conditional upon your actions.

   IMO rationalists find deontology counterintuitive because
   - It's about generalising how social partners subjectively perceive and respond to an interaction (impartially as general relations between persons, i.e. 'objective laws'), rather than predicting which general identities will be externally present.
     Scott Alexander wrote a post about Kant's process-based arguments, where he admitted he thought for a long time that they were silly.
   - It's not based around fixed structures. To reconfigure deontology into a structure-based mould, rationalists used Newcombe's Problem (where a clairvoyant agent represents the intentions of another agent). Christiano jumps covers this under integrity for consequentialists.

4. **Funding effective altruism projects:**
   Project Processes vs Outcomes (excerpt from a grant application I just sent in)

   I want to get more professional at executing on each of the following:

   - Spot a gap as to where the community focuses its thinking and efforts

   - Inquire into considerations on how large a gap is and how useful or tractable a service would be there, in comparison to others

   - Sketch out an inventive solution

- Consult with advisors on a theory of change and onboard any collaborators to make that change happen
- Coordinate efforts to iteratively improve a service and review feedback on whether it actually helps our target users do more good
- Routinise workflows; pay volunteers who were excellent at role; advise managers

At each stage, I want to improve processes for coordinating on making better decisions. Particularly, to couple feedback with funding better:

deserved recognition
actionable feedback ⇅ - ⇅commensurate funding
improved work

What is your project's goal / impact / etc.
To be honest, when I explore an idea with stakeholders for addressing infrastructure gaps, I often struggle to come up with impact scenarios that I wouldn't drastically change a month later.
It can also be dangerous here to make this an official project early on, or to anchor on an end goal for everyone to optimise for. E.g. I aim to test unexplored ideas with collaborators that have the potential to found an organisation around, but my aim shouldn't be 'to incubate x meta-organisation'.

Rather than take up your and my attention explaining why any given wacky-sounding project idea I have could be high-impact down the line, it's more fruitful for me to spend that attention on carefully refining, getting feedback on, and re-prioritising my ideas.

…
Personally, I think my project work could be more robustly evaluated based on
- feedback on the extent to which my past projects enabled or hindered aspiring EAs.
- the care I take in soliciting feedback from users and in passing that on to strategy coaches and evaluating funders.
- how I use feedback from users and advisors to refine and prioritise next services.

Focus in modelling the outside environment:

5. **Forces and mechanisms driving AI developments (vs. agent-specific capabilities/preferences/etc)**

Excerpts from Critch's recent post:

... I'd like you to focus on *multi-agent processes with a robust tendency to play out irrespective of which agents execute which steps in the process.* I'll call such processes Robust Agent-Agnostic Processes. (RAAPs).

A group walking toward a restaurant is a nice example of a RAAP, because it exhibits:

- *Robustness:* If you temporarily distract one of the walkers to wander off, the rest of the group will keep heading toward the restaurant, and the distracted member will take steps to rejoin the group.
- *Agent-agnosticism:* Who's at the front or back of the group might vary considerably during the walk. People at the front will tend to take more responsibility for knowing and choosing what path to take, and people at the back will tend to just follow. Thus, the execution of roles ("leader", "follower") is somewhat agnostic as to which agents execute them.

Interestingly, if all you want to do is get one person in the group not to go to the restaurant, sometimes it's actually easier to achieve that by convincing the entire group not to go there than by convincing just that one person. This example could be extended to lots of situations in which agents have settled on a fragile consensus for action, in which it is strategically easier to motivate a new interpretation of the prior consensus than to pressure one agent to deviate from it.

I think a similar fact may be true about some agent-agnostic processes leading to AI x-risk, in that agent-specific interventions (e.g., aligning or shutting down this or that AI system or company) will not be enough to avert the process, and might even be harder than trying to shift the structure of society as a whole. Moreover, I believe this is true in both "slow take-off" and "fast take-off" AI development scenarios

This is because RAAPs can arise irrespective of the speed of the underlying "host" agents. RAAPs are made more or less likely to arise based on the "structure" of a given interaction.

…
Not calling everything an agent. In this post I think treating RAAPs themselves as agents would introduce more confusion than it's worth, so I'm not going to do it. However, for those who wish to view RAAPs as agents, one could informally define an agent R to be a *RAAP running on agents* A1…An if*:*

...
...

The Production Web as an agent-agnostic process

The first perspective I want to share with these Production Web stories is that there is a *robust agent-agnostic process* lurking in the background of both stories—namely, competitive pressure to produce—which plays a significant background role in both. Stories 1a and 1b differ on when things happen and who does which things, but they both follow a progression from less automation to more, and correspondingly from more human control to less, and eventually from human existence to nonexistence. If you find these stories not-too-hard to envision, it's probably because you find the competitive market forces "lurking" in the background to be not-too-unrealistic.

... If one of the above three Production Web stories plays out in reality, here are two causal attributions that one could make to explain it:

Attribution 1 (agent-focused): humanity was destroyed by the aggregate behavior of numerous agents, no one of which was primarily causally responsible, but each of which played a significant role.

Attribution 2 (agent-agnostic): humanity was destroyed because competitive pressures to increase production resulted in processes that gradually excluded humans from controlling the world, and eventually excluded humans from existing altogether.

The agent-focused and agent-agnostic views are not contradictory, any more than chemistry and biology are contradictory views for describing the human body. Instead, the agent-focused and agent-agnostic views offer complementary abstractions for intervening on the system:

- In the agent-focused view, a natural intervention might be to ensure all of the agents have appropriately strong preferences against human marginalization and extinction.
- In the agent-agnostic view, a natural intervention might be to reduce competitive production pressures to a more tolerable level, and demonstrably ensure the introduction of interaction mechanisms that are more cooperative and less competitive.

Where's the technical existential safety work on agent-agnostic processes?
... I'm concerned that among x-risk-oriented researchers, attention to risks (or solutions) arising from robust agent-agnostic processes are mostly being discovered and promoted by researchers in the humanities and social sciences, while receiving too little technical attention at the level of how to implement AI technologies. In other words, I'm concerned by the near-disjointness of the following two sets of people:

a) researchers who think in technical terms about AI x-risk, and

b) researchers who think in technical terms about agent-agnostic phenomena.

Note that (b) is a large and expanding set. That is, outside the EA / rationality / x-risk meme-bubbles, lots of AI researchers think about agent-agnostic processes. In particular, multi-agent reinforcement learning (MARL) is an increasingly popular research topic, and examines the emergence of group-level phenomena such as alliances, tragedies of the commons, and language. Working in this area presents plenty of opportunities to think about RAAPs.

Remmelt's paraphrase:

Take a future board that buys a particular 'CEO AI service' to ensure their company will be successful. The CEO AI elicits trustees for their inherent categorical preferences, but what they express at any given moment is guided by their recent interactions with influential others (e.g. the need to survive tougher competition by other CEO AIs). A CEO AI that plans company actions based on preferences elicited by board members' preferences at any given point in time, will by default not account for actions bringing into existence processes that actually change the preferences board members state. That is, unless safety-minded AI developers design a management service that accounts for this circuitous dynamic, and boards are self-aware enough to buy the less-profit-optimised service that won't undermine their personal integrity.

Remmelt's interpretations of above excerpts on RAAPs:

- A neat case IMO for how predicting & representing *processes* is complementary (to predicting & representing *structures*) but can lead you to construe and act towards different conclusions.
- Particularly: you can design an AI to optimise for end states based on human-input derived preferences, while it fails to compute how its instrumental actions reinforce dynamics in which certain preferences are expressed.
- Critch appears to intentionally nudge AI x-safety researchers away from construing problems predominantly in a structure-based way (e.g. agent features, states of the environment).
- Nevertheless, he offers readers handholds to convert from process back to structure ('for those who wish to view RAAPs as agents').