# 1 Motivation

**Changes of Heart.** The standard picture of an agent is a system trying to maximize a fixed reward function, or satisfy some fixed set of preferences. However, humans usually either do not always have preferences, or have very poor access to them. Moreover, they often seem to change their fundamental goals, and form new preferences about things they had previously never experienced or understood.

The standard response to this is that really the humans must had have a deeper, higher level goal all along, and that they now have realized that this goal would be better serviced with a different instrumental goal. For rational agents with a fixed reward function do not change their preferences, except insofar as they have totally misunderstood what their actions actually do.

To be clear, this is a perfectly coherent view: humans change their preferences because they are flawed reasoners, and truly rational beings would never do such a thing. Perhaps you were giving money to your kids because you didn't realize that you could have made a bigger impact by donating it to a charity, or perhaps because there is some uncertainty about how effective the charity actually is. And then when you visit the slums of India, you realize that (despite being a kind, positive person being your goal all along). According to this view, morality is an accident, and corresponds to a very deep preference that every person is born with — and any extent to which this picture doesn't seem right, is because you are merely a flawed mortal.

However, it is not a view that feels subjectively correct; the things you care about on a deep level seem to genuinely change: it seems possible to visit a place of poverty and empathize with people to the point of changing priorities, without learning any new facts about the severity of the situation. Moreover it suffers a number of issues on the developmental front: complicated platonic ideals of meta-preferences seem unlikely to be already fully-formed in toddlers, especially given that self-awareness develops at 18 months, and even 5 year-olds stick to simplistic, causal answers to ethical questions.

**AI Safety Concerns**. Using this framework, we have built a substantial theory of how we can use clever optimization techniques to maximize rewards, formulating reinforcement learning (RL). Supervised and unsupervised learning problems can be formulated as restrictions on the general RL problem. A great deal of the problems with learned models are due to overfitting.

We are interested in questions:

1. Can we explain agents' changes of preference without appealing to higher ideals, or by simply stating that humans are flawed? In particular, can we use this theory to explain how values get captured by gamification (grades capturing learning, steps capturing exercise)?
2. The practice of
3. Is it evolutionarily beneficial
4. How does this relate to the problems
5. What implications does this have for designing

, as well as the development of a more general class of reward functions, we must first develop a formal picture of agency, so that we may capture the notion of a "sub-agent", and understand how a reward might arise.

### 1.1 Motivating Examples

## 2 Related Questions and Previous Work

### 2.1 The Orthogonality Hypothesis

## 3 Cartesian (Unembedded) Agents

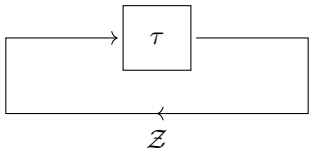
We will develop the discrete case first: an ascription of agency to a nondeterministic process, how it can be extended in the case that the non-determinism comes with probabilities (i.e., is a Markov Process).

### 3.1 General Non-deterministic Processes

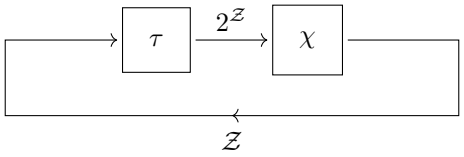
At its very most basic, our input is a space $\mathcal{Z}$, of all possible states that the system could be in, plus a non-deterministic function

$$\tau : \mathcal{Z} \to 2^{\mathcal{Z}}$$

which describes the allowable next states from any given current state. We can also draw this as a *string diagram*, akin to a circuit or process diagram, in which the nodes are functions, and are connected by typed wires which are thought of as carrying data.




This picture would make sense if the system were deterministic (and $\tau$ were a function $\mathcal{Z} \to \mathcal{Z}$), but since the output of $\tau$ is technically different from its input, we need some choice to close the loop like this — like, for example, a function $\chi : 2^{\mathcal{Z}} \to \mathcal{Z}$, which selects one of the possible next states.




This picture, too, is somewhat misleading, since there are many functions from $2^{\mathcal{Z}} \to \mathcal{Z}$ which do not always select a member of the subset in question. To add the appropriate restrictions, we require that $\chi \circ \tau$ is *compatible* with $\tau$ in the following sense:

**Definition 1.** A function $f : \mathcal{Z} \to \mathcal{Z}$ is *compatible* with a non-deterministic function $\tau : \mathcal{Z} \to 2^{\mathcal{Z}}$, and we write $f : \tau$, iff $\forall z \in \mathcal{Z}.f(z) \in \tau(z)$, or equivalently, if

$$\begin{array}{ccccc} \mathcal{Z} \times \mathcal{Z} & \xrightarrow{\tau \times f} & 2^{\mathcal{Z}} \times \mathcal{Z} & \xrightarrow{\mathrm{ev}_{\mathcal{Z},2}} & 2 \\ {\scriptstyle\Delta}\uparrow & & & & \uparrow{\scriptstyle\top} \\ \mathcal{Z} & & \longrightarrow & & * \end{array}$$

**Definition 2.** More generally, $f : \mathcal{Z} \to 2^{\mathcal{Z}}$ is *compatible* with $\tau : \mathcal{Z} \to 2^{\mathcal{Z}}$, and we write $f \subseteq \tau$ iff $\forall z \in \mathcal{Z}.f(z) \subseteq \tau(z)$.

One can see how $\chi$ can be considered an "agent", as it is making the relevant choices. However, this may not be the only way of ascribing agency, and may in fact be far off from what we would normally consider an agent: such an agent may be making choices about very unrelated things in a very large universe. If we interpret the laws of physics in this framework, then this agent is an all-powerful being which makes all choices at every time-step; we want a way of breaking this into smaller pieces.

We want to also (at the very least) allow for a separation between individual choices which are independent from one another, and allow for some subset of the non-determinism to be interpreted without agency.

## 3.2 System Factorization

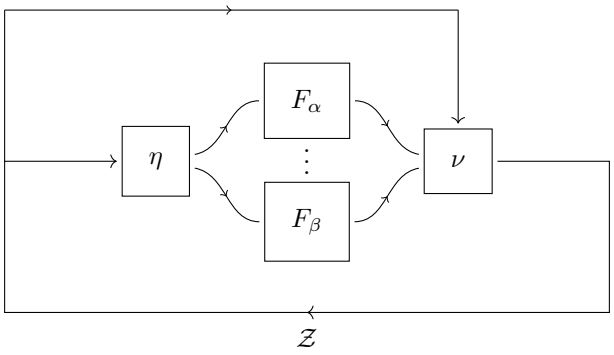
**Definition 3.** An agent factorization of $\mathcal{Z}, \tau : \mathcal{Z} \to 2^{\mathcal{Z}}$ is a tuple $(\mathcal{A}, X, Y, \eta, \nu)$, where

- $\mathcal{A}$ is a collection of agents
- $X = \prod_{a:\mathcal{A}} X_a$ and $Y = \prod_{a:\mathcal{A}} Y_a$ are the input and output spaces, indexed by agent
- $\eta\colon \mathcal{Z} \to \prod_{a:A} X_a$ is a function with components $\eta_\alpha : \mathcal{Z} \to X_\alpha$ for each $\alpha \in \mathcal{A}$, which constructs the input for agent $\alpha$ for a given state of the world
- a function $\nu\colon \mathcal{Z} \times (\prod_{\alpha:\mathcal{A}} Y_\alpha) \to 2^{\mathcal{Z}}$ which uses the previous state of the world and all of the agents' choices to specify a set of possible outcomes

whose composition is compatible with $\tau$, no matter what choices the agents make. More precisely, for any collection of functions $\{F_\alpha\colon X_\alpha \to Y_\alpha\}_{\alpha:\mathcal{A}}$, we have

$$\nu \circ \left[ \mathrm{id}_{\mathcal{Z}} \ \otimes \ \left( \bigotimes_{\alpha:\mathcal{A}} F_\alpha \right) \eta \right] \ \subseteq \ \tau \tag{1}$$

With string diagrams, we have the following picture:




## 3.3 Factorization Properties and Definitions

If, more than simply being compatible with $\tau$, the subset relation in equation (1) is replaced with set equality, then we have exactly captured the nondeterminacy of the system by the choices we expose via the types $X_\alpha$ and $Y_\alpha$. More formally,

**Definition 4.** We will call an agency factorization $(\mathcal{A}, X, Y, \eta, \nu)$ of $(\tau, \mathcal{Z})$ *complete* if

$$\nu \circ \left[ \mathrm{id}_{\mathcal{Z}} \ \otimes \ \left( \bigotimes_{\alpha:\mathcal{A}} F_\alpha \right) \eta \right] \ = \ \tau$$

**Note 1.** *An agent factorization is complete if and only if, for any $z \in \mathcal{Z}$ and $z' \in \tau(z)$, there exists some collection $\{F_\alpha\}$ such that $z' \in \nu(z, \{F_\alpha \eta_\alpha(z)\})$.*

Of course, there's always the trivial factorization, with no agents, trivial $\eta$, each $X_\alpha = Y_\alpha = *$, and $\nu = \tau$, corresponding to the view that everything is deterministic and there are no real choices worth ascribing agency to.

However, we want to focus on factorizations that make sense in terms of agency or influence.

# 4 Embedded Agents

In the previous setting, we considered agents who could have made any choice, because the laws of physics were non-deterministic and loose. In practice, a lot of the things we consider "decisions" are determined hugely by the states of our brain and mental processes we've already picked up. In this case, there is a sense in which there's not really a decision to be made, the choice arises directly from the physics. But there may still be a sense in which we can parse this as a choice — but for this we need a more careful causal model of the world.

In this case, the functions $F_\alpha$ are (at least partially) determined by the state of the world. For now, we will go to the opposite extreme, and characterize the structure when the world is entirely deterministic.

## 4.1 Deterministic Formulation

Once again, let $\mathcal{Z}$ be the state of the world, and now $\tau \colon \mathcal{Z} \to \mathcal{Z}$ is a deterministic function. Just as before, we want to admit descriptions of agency by factorizing $\tau$ into a number of sub-pieces which make decisions – except now the code for running the choices is already determined by some part of the world that is local to the agent.

In order to do this, we will need a way of getting the sub-components of a world state $z \in \mathcal{Z}$, which determine the local state for an agent $\alpha$. This data consists of valid local states $L_\alpha$ is a dependent function

$$\ell \colon \prod_{\alpha : \mathcal{A}} \mathcal{Z} \to L_\alpha \qquad \left[ \text{or} \quad \ell \colon \mathcal{Z} \to \prod_{\alpha : \mathcal{A}} L_\alpha \right]$$

# 5 Sub-Agents

# 6 Evolution

## 6.1 Inner Optimizers

## 6.2 Co-Evolution: Examples and Ties

## 6.3 The Red Queen

# 7 Misalignment and Value Capture

Often, instrumental goals accidentally become final goals, and people become overfit to them. Whereas once a student may have cared about learning, they now optimize for grades instead. One can become fixated on

maximizing steps instead of fitness — on avoiding meat rather than saving animals.

In order for this kind of thing to happen, we postulate that there must be some reason for the internal agent to set up an easier metric to compute and optimize that (it gives gradient info). Note that increasing computational power of an agent, or decreasing the impact of its actions will both have the effect of reducing the pressure for a surrogate reward function to arise.

# 8 Formalism

| | |
|---|---|
| $\mathcal{Z}$ | Space of information in the world |
| $\tau \colon \mathcal{Z} \to \mathcal{Z} \to \Omega$ | Evolution of the world agents in it |

| | | |
|---|---|---|
| $\mathcal{A} : \mathcal{U}$ | | Set of agents |
| $X \colon \prod_{\alpha : A} X_\alpha$ | $X_\alpha$ | Input type for agent $\alpha$ |
| $Y \colon \prod_{\alpha : A} Y_\alpha$ | $Y_\alpha$ | Set of actions for agent $\alpha$ |
| $\epsilon \colon \mathcal{A} \times \mathcal{Z} \to \Omega$ | $\epsilon_\alpha \subseteq \mathcal{Z}$ | Subspace of $Z$ that determines behaviors of $\alpha$ |
| $Q_\alpha \colon \epsilon\alpha \to X_\alpha \to Y_\alpha$ | x | |

# 9 Experimental Plans