This document is intended to be clear, simple, and at a high level. I will use no numbers, prove nothing here, and just make some arguments and provide examples.

# 1 A Very Brief Description

This is an attempt at a model of joint preference and belief revision. The representation is flexible: agents need not always have a perfectly consistent picture of reality, intermediate representations have meanings, and we can capture phenomena such as learning preferences from data and revising them when they're inconsistent — while also disallowing value changes which would be objectionable from an agent's current perspective.

There are two components involved:

1. Representation: An account of an agent's values as distributed across a network, whose nodes are preferences on small sets of alternatives, and whose edges are beliefs about the impact of one choice on another.
2. Dynamics: A formulation of how to revise preferences due to inconsistencies that may arise as beliefs change, and the representation shifts due to new concepts.

# 2 Why This Is Worth Doing

*A few short motivations from several perspectives*

## [Ethos of Agency] The separation between values and tools isn't clean

The distinction between the things that you care about (e.g., utilities, goals, preferences, objectives), and the tools you have to make these things come about (e.g., beliefs, reasons, plans, optimization, inference) plays a large role in technical accounts of agency[1]. Generally the two are considered separately, and also people tend to fix preferences[2]. This is nice because the fixed preferences provide a clean way to evaluate the quality of decisions, and criteria for success are a very important part of science.

Unfortunately, human preferences and beliefs often depend on one another, and in more complex settings the line is much more blurred:

- When you play a video game, are you trying just so you can win? If so, why do you care about winning? It buys you very little socially, and costs time and money. Or does the process of optimizing for this arbitrary goal itself have value? Does this make entertaining yourself fundamentally irrational?
- It is common to have beliefs about preferences: ('I think I like cheese', 'I believe freedom is bad'), and also preferences over beliefs ('Believing true things is good', 'I want to know calculus')
- These can be nested quite deeply without anyone thinking too hard: If I ask if you like cheddar, you now believe that I want to know whether you like cheese — note that this is a belief about a desire for a piece of knowledge about a preference, spanning multiple people.
- On a small scale, we think of a loss function as an optimization objective, and the algorithm (say, stochastic gradient descent) as optimization power. One can use both of these together to form a neural network, which is thought of as being just a tool, with no

[1] Beliefs should be considered optimization power: they you believe that action $A$ will have effect 1, an action $B$ will have effect 2, and a preference between the 1 and 2

[2] In the context of optimization, this corresponds to the practice of writing down an objective and finding extrema; in the context of decision theory, it is finding the best actions to satisfy some preferences

- Complete agents can be used in either way: people, who have desires and experience emotions, can be used as optimization power with an external objective (e.g., hiring employees), but can also as the ultimate source of value (e.g., running an organization because it will help people)

The point is that even we could combine any model of generic optimization, with any model of value, the interactions between them would certainly be woven together, and their dynamics would be intertwined.

This is why we model the dynamics of beliefs and preferences together, and why in some cases we will be able to represent a belief as a preference or vice versa. [3]

## [Modeling Accuracy] Humans change their preferences

When you're born you have no conception of what foods or professions or ideals are good; all you have is your own evolutionarily programmed pleasure and pain. Not only does this feeling get more sophisticated in a way that seems to depend on the environment, but also later on in life, people willingly sacrifice their experiential pleasure for other things they care about. All of this stands even if we didn't see change in the more obvious toy cases: foods, activities, and luxury goods. Such changes do in fact occur, and do not detract from a person's ability to be conceptualized as a coherent agent; often they contribute to a person's character instead.

The standard models do not account for any of this, and for many good reasons — but dynamic preferences are necessary to capture a great deal of human behavior, and are much more important in a world where value is informational and changes quickly. Things like memes, fashion, games, conversations, lofty ideals, and art—things that standard economic theory has had a lot of difficulty ascribing value to—certainly play a bigger economic role than they have in the past, and arguably move faster as well.

This is why we model changes in preferences, and acquisition from only a very limited set of base values. We have some reasons to suspect that in most cases preferences generated this way will be roughly convergent, which if true would (1) provide an additional explanation (beyond empathy and genetics) for why humans end up sharing a lot of values, and (2) allay some concerns about misaligned AIs that are constructed in this way

## [Computational Tractability] Computational Boundedness

The standard picture of decision theory requires keeping around preferences and beliefs about all possible things that are relevant to any decisions you make. To fully describe a general agent's values, then, you need to specify a preference ordering over all possible histories of settings of observable features. This is wildly intractable, and also annoyingly depends on what "possible" means, which is part of the agent's internal beliefs[4]. This is why people in practice restrict the scope of an agent to only a few modeler-chosen variables, assume that they only encounter one kind of decision, specify objectives in a compressed syntactic form.

But even these more tractable restricted approximations we use do not degrade gracefully: in order to make any decisions at all you have to do expected utility calculations (which could be very expensive depending on how clear the impact of your decision on the world is), and there's certainly no clear way of making use of partial computations under time pressure.

By keeping components of preferences around redundantly in many different forms, not only can we immediately re-use them for recent decisions we've made in a pinch, but also likely reduce the complexity of adding new

[3] the categorical interpretation, where each node is a category, the big graph forms a 2-category, and beliefs are value-preserving functors, supports the observation that the distinction between preferences and beliefs isn't a fundamental feature the objects themselves but rather how they're used— here the arrow category here gives us preferences over beliefs. Similarly, quotient objects represented as arrows, just as utilities can be identified with beliefs about utilities.

[4] Or alternatively, if you're going to do the work beforehand as a modeler, has to change when humanity discovers new features of the universe

nodes. This is because we can split the computation into meaningful chunks (one for each preference domain (node) we can connect a decision to), and computing impact on a few nodes of your choice is going to be much faster than computing the total impact on the space of all things that are observable

## [Value Modesty] A second brittleness

Classical AI systems can do complex tasks that even modern ML systems struggle with. However, they have a debilitating flaw: any malformed input a designer has not anticipated causes the entire calculation to be wrong in salvageable ways. It's coded without an outside view of the explicit purpose: there's no possibility that the designer was *wrong* in the algorithm specification, and there's no reasonable inference to make instead even we acknowledged it was possible she was.

Incorporating probabilities and specifying objectives instead of algorithms has helped this problem enormously: by using a ton of redundant, noisy, examples of the algorithm working correctly, we can be reasonably resistant to malformed input. Unfortunately, these systems are not perfect either: they are often biased and sometimes cheat. We face a different kind of brittleness here: what if we forget about a second order effect and slightly mis-specify an objective function? What if we run it on a different input distribution? ML systems won't crash, but they will confidently display wrong answers. The system doesn't think there's any possibility that its *objective is wrong*, and even if it acknowledged this possibility, it's not clear there's anything to be done at this point.

In analogy to the solution that statistical ML provided to the problem of brittle classical AI, the solution would seem to be a noisy, redundant encoding of the objective function in conjunction with a meta-objective: in our case, consistency. The diffusion of preferences (which we will also refer to as value capture) can also be thought of as a source of uncertainty about your values even if you didn't start with any: if there are multiple objectives consistent with everything you've seen, you acquire a preference for both of them.

Of course, humans are also often uncertain about their values (and experience value capture) in addition to their beliefs and the appropriate choice of actions.

## [Theoretical Unity] Reductions to other theories and algorithms

Although the aim in the rest of this document is to ignore generality and aim for descriptively simple, the generality is a genuine mathematical motivation for this theory: it can be viewed as a number of different things. I will briefly mention some of them that I'm excited about here, and then focus on simple things for the rest of the document.

- Our model reduces to standard expected utility calculations in the degenerate case where the agent has preferences over possible histories of worlds, the representation of a world never changes, and the agent is cognitively unbounded
- It has a categorical interpretation which has some features I'd like to explore: in particular, meta-preferences seem to be related to higher order structure, a total utility function is like a limit, a total probability function is like a co-limit, preference-preserving beliefs are functorial, and the nodes already form a 2-category, whose objects are themselves enriched flat categories.
- There also a natural interpretation of this as an artificial neural network. If the underlying graph is a DAG with one sink, then it is a feed-forward network for calculating expected utility. In most interesting cases, it will have recurrent connections and no special output.
- For agents that have priors, the beliefs encoded in this way can be converted to a Baysian Network with a simple transformation
- The preference propagation is a bit like broadcasting on a network (of physical machines) to compute path lengths. This can be done Dijkstra / Distributed Bellman Ford [5] and can be computed by taking

[5] depending on whether we take right or left powers of matices

iterative powers of a giant matrix over the tropical semi-ring. Funnily enough, this is exactly how you compute the transitive closure of preference matrices on a small scale (except with a different semi-ring), which lends more credibility to the higher order categorical interpretation.

The possibility of bridging these many disparate fields makes the abstraction feel much more universal, and the ability to think about preference change in any of these terms could make it easy to identify analogs of non-trivial facts in other fields.

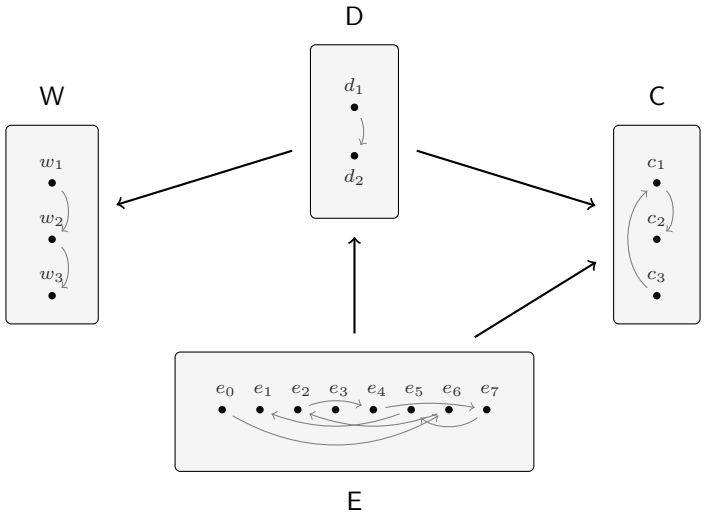
# 3 Longer Informal Description and Intuition

## 3.1 Representation

We will consider preferences on small sets of alternatives (nodes), together with a graph of beliefs, about how one preference impacts another. For example, the nodes could include: the set of ice cream flavors, a collection of memorable experiences, freedom and control, and anything else that could conceivably be considered an exhaustive set of alternatives for a choice. On these sets, we require preferences of some form; while I would like to ultimately propose that they be thought of as semiring matrices, we will stick to utilities for each alternative in the examples here.

In addition, we also need beliefs about the impact of $A$ on $B$. While we do not want to commit to one of these in general yet, conditional probabilities will do the trick to explain the examples which follow. A belief about how $A$ impacts $B$ then, for the remainder of this document, is a family of probability distributions over $B$, one for each alternative $a \in A$ — an object which looks and acts like the conditional distribution $\Pr(B \mid A)$, with which it will intentionally be confused.

The structure of a representation with both pieces might look something like this:

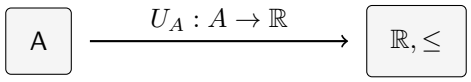



There are alternatives in each of the domains, preferences among the alternatives (we draw $w_1 \to w_2$ to communicate that $w_1 \preccurlyeq w_2$, i.e., that $w_2$ has higher utility than $w_1$), and links (conditional probability distributions, whose values are not drawn here) between them.

### 3.1.1 Representing Utilities, Probabilities

While I will explain more details of the correspondence between this presentation and other standard ones in section 6.1, the examples may be more informative if we can see how utilities and credence enter the picture.

Suppose the only thing in my model is a single variable $A$, which can take on values $a_1, a_2, \ldots$. A preference on the alternatives often takes the form of an order, and if the order is complete and transitive, then it can be represented as a utility function. We can capture this explicitly with our model: think of utility as a separate preference domain $\mathsf{U} \cong (\mathbb{R}, \leq)$, whose elements are the real numbers, with a preference given by the usual ordering. Now a choice of $A$ has an impact on what happens in $\mathsf{U}$, and if we know that each choice of $A$ gives us a deterministic utility, then this impact is actually just a function, and in particular can be represented as a degenerate probability distribution $\Pr(\mathsf{U} \mid A)$, which is just a function from $A$ to $\mathbb{R}$.



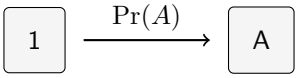

In this sense, we can think of utility functions as a way of representing preferences on $A$ as a belief about how $A$ impacts a standard universal preference domain of "goodness". Similarly, if we have multiple variables which impact one another, we can just take the product of all of their variables, and consider a conditional probability distribution from this to the utility preference domain $\mathsf{U}$.

[**todo**: *Move this so the section isn't as long, this is starting to be out of the critical path*]

So why is the standard to use the real numbers instead of some other ordered set? The biggest reason is that uncertainties need to trade off against value, and we use probabilities to track uncertainty. For example, suppose I think $a_1 \preccurlyeq a_2 \preccurlyeq a_3$. This is enough information to make choices between subsets of the values of $a$, but is not enough to know whether I prefer an $\alpha$ chance at $a_1$ and $(1-\alpha)$ chance at $a_3$, over a sure outcome of $a_2$. This is unsurprising — after all, bets are different things altogether. In fact, we can also put the space of bets into our diagram to clarify what's going on here:


A choice of bets on $A$ (above represented $B[A]$) clearly impact $A$, and in this case our formulation is particularly simple: the conditional probability $\Pr(A \mid B[A])$ is just the bet itself, i.e., $\Pr(A = a \mid B[a] = \vec{p}) = \vec{p}_a$. But if we want the composition of these two links to be equal to the utility on bets (i.e., we want it to be consistent, as discussed in the dynamics section), then we have to be computing expected utility.

On the one hand, we can use these links to describe utility, ignoring the probability distribution over the output, and making full use of the dependence on input. We can also do the exact opposite, and ignore the dependence on input, making full use of the distribution as output. This can be used to give us a prior without any parameters, which is represented graphically as an arrow from the preference domain with only one object, denoted $\mathsf{1}$.




These two are dual in some sense. Intuitively, one of these choices is choice made by the agent, and the other one by the environment. A utility function uses only the dependence on the input, so it can be seen as a vector $u : A \to \mathbb{R}$; a probability distribution makes use only of the distribution over outputs, and can be represented as a vector. If the links are thought of as stochastic matrices, then one is the transpose of the other.

In this way, we can think of passive expected utility calculations, across the values of a variable $X$, as a factorization of $\mathsf{1} \to \mathsf{U}$ through the alternatives of $X$, i.e., $\mathsf{1} \to \mathsf{X} \to \mathsf{U}$.

### 3.1.2 Benefits of this Representation

It allows us:

- to represent inconsistent and redundant preferences

- to introduce new choices, and to modify or delete preference information without needing to globally re-compile a joint preference
- to emulate other preference descriptions (see above, and section 6.1)
- to store preferences about common choices so they can be re-used and nudged without fully recomputing everything for each decision

The most important reason for representing preferences this way is to facilitate preference change, described in a bit more detail below.

## 3.2 Dynamics

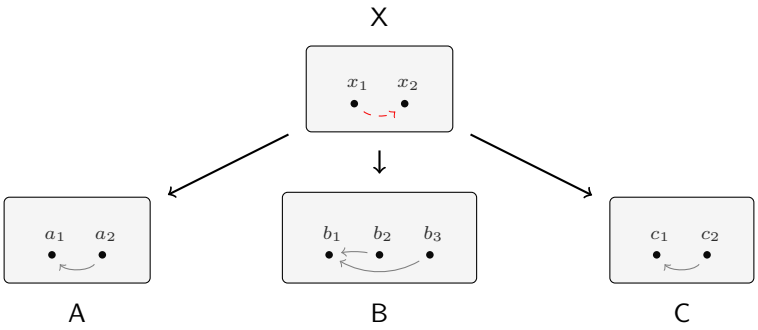
Given a network representation of preferences, we can compute some measure of inconsistency, and then try to move preferences (or even beliefs) around so as to reduce the total inconsistency. This simple procedure has a number of interesting consequences, which we will go through in this section. Later, when we further specify the way this can be done, we will be able to explain several additional phenomena.

### 3.2.1 Value Capture

Suppose you care about

### 3.2.2 Deductive Preference Formation

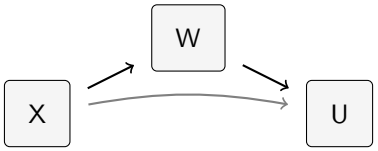
Suppose you're trying to decide what your preference for $X$ is: a standard procedure that humans rely on is the compilation of a (weighted) list of pros and cons. In other words, you examine the impact of a decision in $X$ on other things you care about, say variables $A, B, C \ldots$. Maybe $x_1$ makes $a_1$ more likely, which is a good thing, also $b_3$ and $c_2$ which are both bad. On the other hand, maybe $x_2$ doesn't impact $A$ as much, but makes $b_1$ and $c_1$ more likely, which is ideal, making $x_2$ better than $x_1$. Now, we have external reason to prefer $x_2$ to $x_1$ — which means that there's a preference conflict if we were initially indifferent between the alternatives in $X$! By revising our preferences in $X$ to reflect $x_2 \succcurlyeq x_1$, we can remove the conflict.




Note what we have done: we have a link for how a choice in $X$ impacts downstream things, and so existing preferences flow backwards against the flow of causality, to form a preference on $X$. This is the standard, deductive way of making decisions.

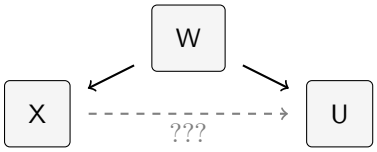
To see more clearly why preferences form *backwards* along links, consider the diagram before, except where we represent preferences are represented by utility functions. If we have a conditional probability $\Pr(W \mid X)$ and a utility function $U_W = \Pr(U \mid W)$, then we can simply compose them to get a utility on $X$. Note that there is not an easy way to combine a utility function on $W$ and a conditional probability that goes the other direction (see the diagram below).

*Easy to compose (deductive formation):*



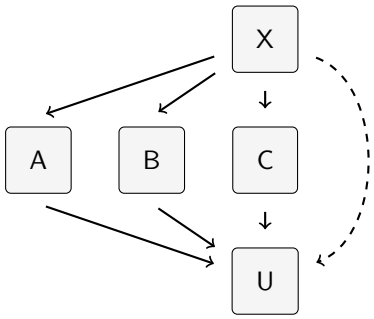

*Existing preference $W \to U$ can be extended naturally to $X$ by precomposition, backwards along $X \to W$*

*Hard to compose (inductive formation):*




*Getting a utility on $X$ via $W$ requires the inversion of an arrow, or some other additional information*

One might wonder how this plays out if we have multiple variables instead of just one. In our earlier example, each of the three variables $A, B$ and $C$ has an additive component of the utility function, which can be visually identified as the three disjoint paths from X to U.




Now, our initial indifference between the alternatives in $X$ conflicts with the rest of the diagram, i.e., the other paths from $X$ to $U$. We can resolve this conflict by setting our preference on $X$ to be the sum of the other three components.

This example shows that we don't need to make all diagrams commute; for consistency we rather require just that a link between $A$ and $B$ is the (weighted) sum of other paths from $A$ to $B$. Inconsistency therefore cannot be determined locally, because we need to see all of the paths.

# 4 Insufficiency of Classical Model: Examples

## 4.1 Framing Problems

Since inconsistency is driving preference changes in this formulation, we will start with the case where clearly this is what's going on. Suppose you prefer $a_1$ over $a_2$ (for variable $A$) and $b_2$ over $b_1$ (for variable $B$). Later, you come to believe that a choice of $a_1$ is really logically equivalent to $b_1$, and $a_2$ to $b_2$; in fact, $A$ and $B$ were the same variable. This happens all the time for humans, if $A$ and $B$ were given different descriptions. For example,

> *You believe that the rich should not get a larger federal tax exemption (than the poor do) for having children. At the same time, you believe that if the default number of children were 2, and people paid a surcharge to the government for foregoing them, that the rich should pay a larger surcharge (than the poor do) for this. However, the default number of children is not a feature of the world, and the same amount of money changes hands in both cases; the two preferences are logically in conflict.*

This creates a conflict— it is impossible to have all four of

$$\Big\{\text{consistency}, \qquad a_1 \succcurlyeq a_2, \qquad b_1 \preccurlyeq b_2, \qquad \text{the belief that } (a_1 \equiv b_1) \wedge (a_2 \equiv b_2)\Big\}$$

Before we go any further, notice that there's not even an obvious, natural representation of this conundrum in terms of classical decision theory: if utilities are over outcomes, and the two descriptions pick out the same

set of outcomes (without changing their relative probabilities), then you will never be in a situation like this. Let's try to salvage this from the classical point of view in a few ways.

#### 4.1.1

One possible patch is to have utilities over *descriptions* of worlds, dependent on more than just the worlds they pick out. Now, this "conflict" from before is not problematic: you prefer $a_1$ to $a_2$, $b_2$ over $b_1$.

We can use this to represent the state of the world, but now:

- Computation of expected utility now requires priors over descriptions of worlds, an object which is even more astronomically complex than a prior over worlds.
- Crucially, this account does not denounce, let alone provide any mechanism for resolving the disagreement, and therefore we sacrifice all of the standard rationality guarantees, such as resistance to dutch booking[6]. Decisions consistent with this picture may have nothing to do with the world may be entirely based on the number of parentheses in the formula.

Clearly, this is a flimsy model as it stands, but it's not clear how to fix it. If we add independence of description as an axiom, we're back to our original representation problem.

#### 4.1.2

Another tactic we compatible with the classical decision theory perspective[7], this time much closer to our own formulation, is to define utility by additively separable components, each arising from a variable (we have once again implicitly given ourselves utilities over descriptions of the world), relying on belief revision to effectively make some worlds impossible.

The set of possible worlds which we have utilities over is once again all possible assignments to variables, and hence includes "impossible possible worlds", such as the one where $a_1$ and $b_2$ are simultaneously true. By belief updating, the likelihood of these worlds can be driven to zero, and when we say "$a_1 \succcurlyeq a_2$", we really mean that the variable $A$ has an additively separable component of our total utility function (resp. for $B$), and it is merely an unfortunate fact of life that good things are attached to bad ones. If the two preferences carry the same weight, then as the possibility of achieving impossible worlds vanish, the two given preferences annihilate one another and we are left indifferent, as far as the total utility function is concerned.

This seems in many ways better, because the tension is resolved, it can happen continuously, the agent is only vulnerable to dutch booking as long as its beliefs are inaccurate, and we get a temporal reason for why the agent's choices change. In other ways, it is less satisfying:

- In order to define a classical agent, a modeler must specify a utility function, and in this case means that each additively separable component must have been there in the beginning. In particular, this means the modeler decides on the set of variables[8], and so agents can only be in this situation insofar as the modeler has given them representational redundancy: it is impossible for an agent to experience conflict in this way if the variables in fact determine unique states of the world.

---

[6] If you have utilities $u_1, u_2, v_1, v_2$ for $a_1, a_2, b_1, b_2$, respectively, then you would be willing to pay $u_1 + v_2$ for a bet of $a_1$ or $b_2$ (which has probability 1), and $u_2 + v_1$ for a bet of $a_2$ or $b_1$, which again is probability 1. Since the difference in utility differs between the $a$ and $b$ descriptions, $u_1 - u_2 \neq v_1 - v_2$, and so $u_1 + v_2 \neq u_2 + v_1$, and a bookie can make unbounded money off of you by selling you one bet and buying the other.

[7] the model described here actually corresponds to an influence diagram with two decision and two value nodes

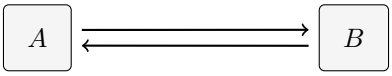
[8] If instead the modeler only described a way of generating these utilities, so that the set of variables could change, then there must be some additional structure to make these decisions, making the picture immediately very non-classical — classical utilities do not depend on an agent's beliefs or other mental features

- There can never be a time when the agent holds both the contradictory preferences, and also the belief that they are contradictory, even if belief updates happen slowly.
- You can never truly change your opinion on either issue: you will always want $a_1$ over $a_2$ and $b_2$ over $b_1$ — even if you now see how they are exactly the same decision.
- Because we define a fixed utility function, the two conflicting components on the same concept could not have been the result of some dynamics mechanism; they were in fact put there (explicitly) by the modeler.

More briefly, our complaint is that the components of the utility function didn't actually change, and that the set of variables is necessarily fixed.

#### 4.1.3 With Our Preference Network Model

We propose a slight modification of the previous model, where preferences actually flow along the beliefs that link them: In our case, this means we consider what it would look like to make a decision about variable $B$ looking only at what we care about in $A$, and vice versa. Since the correspondence is a logical one, the "impact" works both ways, so for us this means we'll use two beliefs, to distinguish it from the case where one causes the other.




To keep this as simple as possible, let's consider the case where the beliefs linking the two domains form instantly, and consists of perfect correspondences, degenerate probability distributions with all of the mass on one point:

$$\Pr(B \mid A) = \begin{cases} a_1 \mapsto b_1 \\ a_2 \mapsto b_2 \end{cases} \qquad \Pr(A \mid B) = \begin{cases} b_1 \mapsto a_1 \\ b_2 \mapsto a_2 \end{cases}$$

With this, our preference $(a_1 \succcurlyeq a_2)$ has an image of $(b_1 \succcurlyeq b_2)$ under the link $B \to A$, which conflicts with our preference on $B$; similarly, $(b_2 \succcurlyeq b_1)$ results in a preference $(a_2 \succcurlyeq a_1)$ which conflicts with our preference on $A$. Changing the belief or either preference can resolve the conflict locally, but as we saw earlier, the actual inconsistency depends on the rest of the network.

### 4.2 Learning from Experience

We can use the same technique of reducing conflict to model preferences learned from data. Let $\mathsf{Exp}$ be a preference domain representing all of your experiences (which you have preferences/utilities over), and $\mathsf{X}$ be some domain consisting of all possible combinations of features (e.g., what food you ate, who you were with, whether you got work done, etc. ). Suppose further that each experience had a utility, and also a setting of the features (we can model even the features being fuzzy by making use of the probability distribution half of our links— maybe you forgot what kind of ice cream you were eating).

#### 4.2.1 Classical Models

The standard picture of preferences does not prescribe change, and so learning is generally outside of the scope. Still, there's another reading which makes sense more-or-less classically:

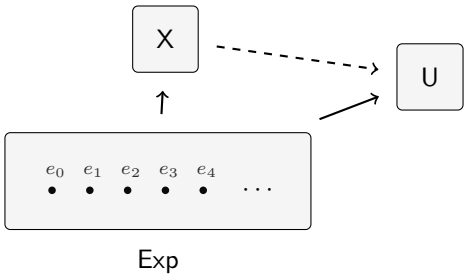
You have some fixed preferences, you just don't know what they are. As you get more samples, you uncover what your preferences must have been all along. The features do not provide any value in and of themselves— they only provide some information about the total state of the world. The experiences are true samples of your fixed utility function.

This can be made to work, but is not a particularly satisfying account:

- Economics outsources the procedure for how to do this to a completely different discipline.
- Humans seem to develop preferences because of experiences, in addition to merely uncovering them — whether or not you have food poisoning the first 10 times you eat mangos, for instance, will probably color your perception of the taste even if you don't get poisoning after this.
- The classical picture also posits that people make decisions according to their preferences, but this is not possible if you don't know what they are — and if it were the case that you were unknowingly making decisions this way, then you could easily access any preferences just by making hypothetical decisions.

### 4.2.2 Our Description

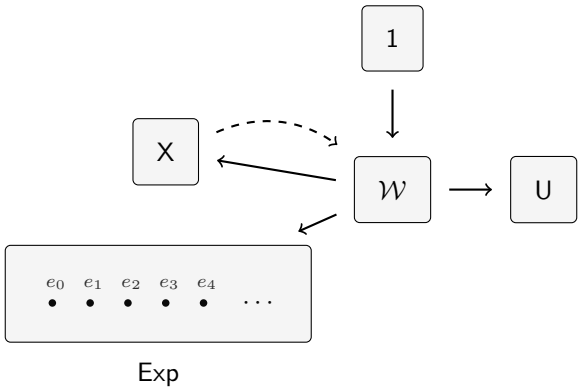
We can succinctly capture the entire setup in this diagram:




To avoid conflict, we would need it to be the case that $\mathsf{Exp} \to \mathsf{X} \to \mathsf{U}$ is the same as $\mathsf{Exp} \to \mathsf{U}$. There are many methods of doing this — we can guess and check, do gradient descent on parameters, set up a fancy network encoding some similarity metrics and priors, ... but in each case, we're just reducing inconsistency, same as before.

It may be worth noting that kernel methods do this in a particular way: they compute a compressed approximation of the inversion of the conditional distribution $\mathsf{Exp} \to \mathsf{X}$. Rather than storing the features for each experience, we can describe features as weighted sums of the experiences they most resemble. Normalizing this weighted sum, we get an arrow $\mathsf{X} \to \mathsf{Exp}$. Having done this approximation, we can then compute utilities for $\mathsf{X}$ by precomposition as in the previous section. Also, computing the full pseudoinverse of $\Pr(\mathsf{X} \mid \mathsf{Exp})$, regarded as a matrix, is just a computationally expensive way of computing a least squares fit on $U$.

The classical picture described above corresponds to this diagram:

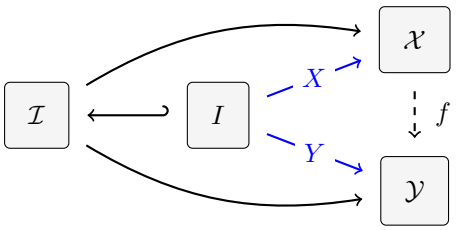



Here $\mathcal{W}$ is the set of all possible worlds, which has a probability measure and utility on it (to the right and above). Really that's all we need, but the world determines which experience you're having and what features as well. You can use the prior over worlds and Bayes' rule to invert the arrow $\mathcal{W} \to \mathcal{X}$ if you really want it, but in any case all you're doing is updating a belief about how $X$ impacts the world so you can compute expected utility better—-you probably don't actually have preferences over $X$, unless you've been guaranteed that your utility function additively separates over some component of it.

#### 4.2.3 General supervised learning as reduction of inconsistency

The picture we've drawn is actually an instance of a fully generic as a description of a supervised learning problem. In the standard formulation, we're given samples $(X, Y) = \{(x_i, y_i)\}_i$, $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the feature space, $\mathcal{Y}$ is the label space, there is some true oracle function $\hat{f} : \mathcal{X} \to \Delta(\mathcal{Y})$, and therefore $(X, Y)$ is the training set, with $Y \sim \hat{f}(X)$. The problem is then to infer a function $f \approxeq \hat{f} : \mathcal{X} \to \Delta(\mathcal{Y})$. The naive solution is to minimize the training error on the samples, but this leads to overfitting, partly because this implicitly involves thinking of a sample $(x, y)$ as representing "the true value of $\mathcal{Y}$ when $\mathcal{X} = x$ is $y$" — when in reality there is some noise: both $x$ and $y$ do not represent a probability distribution of uncertainty, but rather concrete sampled values.

We can describe the same scenario equivalently (and arguably more clearly) in our model. If $I$ is the index set that $i$ comes from, then we've been given a function $X : I \to \mathcal{X}$, $Y : I \to \mathcal{Y}$, and learning function $f : \mathcal{X} \to \Delta(\mathcal{Y})$ such that there's minimal conflict corresponds to minimizing training error (inner triangle in the diagram below)




On the other hand, if $I \subseteq \mathcal{I}$, where $\mathcal{I}$ is the set of all possible possible states of the relevant world, then we can think of minimizing generalization error as reducing the inconsistency of the outer diagram[9]: we want it to be the case that $f$ is approximately correct in all cases, not just the ones indexed by our sample $I$. Bayesian methods make the additional assumption that we have a prior $1 \to \mathcal{I}$, which can be used with Bayes' rule to invert arrows.

This also makes the source of overfitting mentioned above clearer: the training error will also be minimal when we find the best fit, if we were provided $X$ and $Y$ as conditional distributions rather than samples.

## 5 More Examples

[**todo**: *I have a list of some of these, commented out because I haven't put in the work to explain any of them. They're just place holders for me as I worked things out; I'll uncomment them as I get time to explain them, but they're low priority, merely serving to address concerns I had but didn't articulate*]

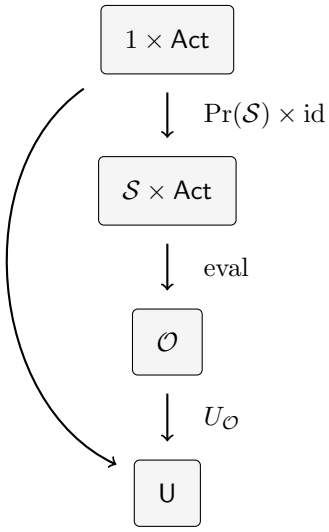
[9] which is equivalent to minimizing the inconsistency of the outer triangle, because the top and bottom ones both commute because $I \hookrightarrow \mathcal{I}$ is an inclusion

# 6 Ties To Existing Models

## 6.1 Emulating Other Structures

### 6.1.1 Savage

Savage's theory can also be represented in this context. If you have a preference order on $\mathsf{Act}$, which is the set of all maps from states $\mathcal{S}$ to outcomes $\mathcal{O}$, which obeys Savage's postulates P1-P7, then it can be factored into a probability distribution $\Pr(\mathcal{S})$ and a utility function $U_{\mathcal{O}}$, as shown on the right.

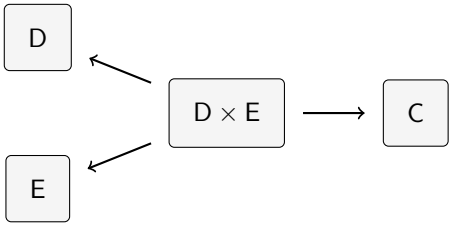



## 6.2 Important Differences

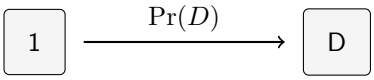
### 6.2.1 Departure from BNs and CP Nets

Although this representation looks like a Bayseian Network, and in fact any BN can be encoded in this way, the two formulations differ at face value in several important ways:

1. We encode preference information in each domains,
2. The existence of two arrows into $\mathsf{C}$ in the diagram above does not mean that there is a distribution $\Pr(\mathsf{C} \mid \mathsf{D}, \mathsf{E})$, but rather two different ones $\Pr(\mathsf{C} \mid \mathsf{E})$ and $\Pr(\mathsf{C} \mid \mathsf{D})$. It is possible to recover this meaning if we are allowed to introduce new product nodes:




3. Similarly, a node without any parents has a probability distribution on it in a BN; this is not the case for us (think of this as the zero-arity product, making this a special case of the above). Once again, this can be represented by explicitly adding an edge, from the singleton preference domain:




As a result, a model like this does not always represent a factorization of the joint distribution on the product of the variables. In some sense, it does represent a factorization of joint preferences on the variables, but not in a way analogous to the BN approach (this is what CP Nets do).

While it is true that I have implicitly defined a utility function,

### 6.2.2 Static description in terms of Influence Diagrams

Our rounded nodes function somewhat like a hybrid of the three nodes used in influence diagrams: aleatory variables (circles), decisions (squares), and values (octagons). [todo: ]