Communication & Trust

Abram Demski

September 16, 2025

Abstract

Yudkowsky suggested the criterion of *reflective consistency* for decision theories (roughly: does a decision theory choose itself?) [Yud10]. Dai proposed Updateless Decision Theory (UDT) as a response to Yudkowsky's ideas [Dai09]. [DHR25] offered the first published proofs of reflective consistency for UDT. However, those results were not entirely satisfying, due to their reliance on strong assumptions. The current work offers a new attempt, inspired by Critch's notion of agent boundaries [Cri22] as well as Garrabrant's work on Cartesian Frames [GHLW21] and Finite Factored Sets [Gar21]. The approach here uses *communication between agent-moments* as a "release valve" for pressures which could otherwise lead to self-modification.

1 Introduction

Self-trust is an important safety property for agentic AI. Without such trust, AI systems have an incentive to modify themselves or create successor agents, which could undermine other safety properties. A better understanding of trust could also contribute to safety in other ways; see [DHR25] for further details.

The current paper, like [DHR25], analyzes conditions under which multiple instances of an agent (across space and/or time) can justifiably trust each other. Trust is operationalized as non-interference: given the opportunity to modify how an instance makes decisions, a preference to do so indicates a lack of trust.

Unlike [DHR25], the current paper focuses on *communication* as a means of creating trust. Without communication, coordination problems can create a lack of trust even between agents with shared goals and beliefs. This resulted in overly strong coordination assumptions for previous results. The current work makes coordination assumptions *only with respect to communication itself*, assuming enough for the agent-instances to have a shared communication protocol. This is used to overcome any other coordination problems which could otherwise break trust.[1]

Section 2 will provide further historical context for the ideas to be presented here, by contrasting the notion of "trust" used here with the common notion of *dynamic consistency*. Section 3 will provide further motivation for the current approach by way of example decision problems.

Section 4 will begin the formal development of these ideas by introducing important notation and mathematical terminology. This includes a notion of factorization inspired by (but distinct from) Garrabrant's work on Finite Factored Sets [Gar21]. Section 5 will apply these mathematical tools to model agents, inspired by (but distinct from) Garrabrant's Cartesian Frames [GHLW21] and Critch's work on agent boundaries [Cri22]. Section 6 elaborates this model to deal with multiple instances of an agent and communication between those instances. Section 7 proves the main result on the avoidance of self-modification. Section 8 deals with the question of whether agent-instances will follow advice that is communicated to them by other instances. Section 9 concludes with a discussion of the significance of the results and future work.

Throughout the paper, I will use "I/my" to take personal responsibility for decisions/thoughts/etc (EG "I will call this variable X ..."), and "we/our" to invite the reader along (EG "With this technique, we can ...").

[1] This strategy owes a significant debt to discussions with Scott Garrabrant, although the strategy he was advocating in those discussions differs considerably from my strategy here.

2 Trust vs. Dynamic Consistency

Suppose that an agent has, by virtue of its constitution (biological or synthetic), a specific decision rule. So long as the agent is functioning normally, this decision rule determines how all instances of the agent make decisions. However, the environment might provide some instances of the agent with opportunities to interfere with other instances (or with themselves)[2], creating a circumstance where an instance is *not* operating normally, and can make decisions which do not conform to the decision rule. Some agents may prefer to interfere with themselves in such a way. The current work formalizes trust as the absence of such a preference.[3]

This closely resembles the concept of *dynamic consistency* which is familiar to both economists and decision theorists [Str55, Pol68, Mac89, FLO02].[4] An agent is dynamically consistent if, *told of its future actions*, it would endorse them.[5] A dynamic inconsistency of this sort implies that an agent would choose to interfere with its future decisions if (a) it had the opportunity to do so, and (b) it could foresee those future decisions precisely.

Yudkowsky contrasts dynamic consistency with *reflective consistency*:[6]

> I wish to generalize the notion of *dynamic consistency* to the notion of *reflective consistency*. A decision algorithm is *reflectively inconsistent* whenever an agent using that algorithm wishes she possessed a different decision algorithm. Imagine that a decision agent possesses the ability to choose among decision algorithms—perhaps she is a self-modifying Artificial Intelligence with the ability to rewrite her source code, or more mundanely a human pondering different philosophies of decision. ([Yud10])

Yudkowsky doesn't offer a precise definition of reflective consistency in that work.[7] However, his later work on the concept of *tiling agents* articulates the *Vingean principle*:

> An agent building a successor (equivalently: a self-modifying agent creating the next generation of its code) should not need to know the successor's exact actions and thoughts in advance. ([YH13])

In the current work, I interpret the Vingean principle as follows: we are not allowed to assume an agent can perfectly predict the strategy of all of its instances. This bars predicting the precise actions of instances (and also bars precise calculation of mixed-strategy equilibria).[8]

In my experience, those influenced by Yudkowsy's ideas use 'tiling' and 'reflective consistency' interchangeably. Both terms are, in my opinion, best understood as what the current work calls 'self-trust': a variant of dynamic consistency where one does not assume that the agent can foresee the decisions of all its instances. Instead, an agent must reason about itself abstractly, establishing trust in its other instances to do the right thing based on shared goals and properties of the shared decision procedure. This better reflects the problems of self-trust in the real world, since realistic agents cannot precisely plan everything out ahead of time.[9]

[2] It may be sensible to assume an instance cannot modify itself, as this would seem to require time-travel. It might further be sensible to assume a temporal partial order, so that later instances can only be influenced by strictly earlier instances. However, the present work avoids such assumptions. This is done to minimize unecessary assumptions, as well as to respect the spirit of Updateless Decision Theory, which is not supposed to rely on any notion of time or causality.

[3] I do not intend to assume that an agent can perfectly identify instances of itself within the environment. What is important is that the agent reasons in such a way as to imagine that there may be instances, and may plan (using its uncertain understanding of the world) to interfere with those instances.

[4] Dynamic consistency is sometimes alternatively called time consistency or intertemporal consistency.

[5] More precisely, dynamic consistency is often defined as consistency between the plans an agent would make ahead of time with the decisions which it would make in the moment. However, this amounts to the same thing.

[6] Pettigrew et al [PCKng] examine a very similar concept which they term *self-recommending* (calling its negation *self-undermining*).

[7] "I have never seen a formal framework for computing the relative expected utility of different abstract decision algorithms, and until someone invents such, arguments about reflective inconsistency will remain less formal than analyses of dynamic inconsistency." [Yud10]

[8] I do not intend this as the only or ultimate interpretation of the Vingean principle. In [YH13], the authors state "For our purposes we cash out the Vingean principle as follows: *In the parent's reasoning, the offspring's actions should only appear inside quantifiers*."

[9] A possible distinction between the current work's notion of 'trust' and reflective consistency, on the one hand, and dynamic consistency and tiling on the other hand: the second pair are described in relation to time, whereas the first two deal with 'instances' which may be across time or otherwise; however, I see this as a less important distinction.

Dynamic consistency is sometimes treated as a fundamental rationality constraint; for example, Dutch Book [Vin22] and Money Pump arguments [Gus22] can be interpreted as dynamic inconsistency arguments (illustrating a sequence of decisions which an agent would endorse individually, but would not endorse as a plan chosen all at once). I read Yudkowsky as intending to suggest that reflective consistency is equally or more fundamental (and I am inclined to agree). However, Yudkowsky cautions:

> Therefore I cannot say: If there exists *any* dilemma that would render an agent reflectively inconsistent, that agent is irrational. The criterion is definitely too broad. Perhaps a superintelligence says: "Change your algorithm to alphabetization or I'll wipe out your entire species." [...] To make *reflective inconsistency* an interesting criterion of irrationality, we have to *restrict* the range of dilemmas considered fair. I will say that I consider a dilemma "fair," if when an agent underperforms other agents on the dilemma, I consider this to speak poorly of that agent's rationality. ([YH13])

Yudkowsky goes on to discuss how different notions of fairness will lead one to endorse different decision theories. Yudkowsky endorses a specific notion of fairness, which he calls *decision-determination*.

> An expected utility maximizer can succeed even on problems designed for the convenience of alphabetizers, if the expected utility maximizer knows enough to calculate that the alphabetically first decision has maximum expected utility, *and if the problem structure is such that all agents who make the same decision receive the same payoff regardless of which algorithm produced the decision*. This last requirement is the critical one; I will call it *decision-determination*. ([Yud10])

The main contributions of the current work are to offer a particular formalization of Yudkowsky's notion of fairness, and prove a result analyzing the reflective consistency (self-trust) of Wei Dai's Updateless Decision Theory.

3 Motivating Examples

In response to some of Yudkowsky's ideas about decision theory, Wei Dai proposed Updateless Decision Theory (UDT) [Dai09]. Its decision rule is as follows:

$$\pi(o) = \arg\max_{a \in A} \mathbb{E}(U|\pi(o) = a)$$

Here, π is the agent's policy; that is, a function taking observations $o \in O$ and outputting actions $a \in A$. U is the agent's global utility function, and $\mathbb{E}$ takes the expectation in terms of the agent's prior. (This notation will be revised and elaborated later.) This decision procedure achieved self-trust in many examples of interest where other decision procedures failed.

Dai later noticed examples where the above decision rule fails to achieve self-trust [Dai10]. These examples are *coordination problems*, where multiple agent-instances must take coordinated action to achieve a desired outcome. For example:

Example 1 (Coordinated Buttons problem)**.** [10] You are about to be copied. You will be put in a red room, while your copy will be put in a green room. In both rooms, there are two buttons: one labeled \$5, and another labeled \$10. Each copy can press one button. If you both press matching buttons, then you both receive the amount of money written on that button. Otherwise, you receive no money. Before being copied, you have an opportunity to take a pill which will override your thinking for the duration of the problem, giving you the overwhelming reflex to press buttons labeled \$10.

Intuitively, the correct answer is to press the \$10 button, and be indifferent about taking the pill (because you'll do the same thing whether you take the pill or not). Unfortunately, UDT's answer depends on its prior! If you have a sufficiently high prior expectation that you'll press the \$5 button, then both red-room self and green-room self will prefer to press \$5.[11] If you anticipate this, then you will prefer to take the pill before being copied, reflecting a lack of self-trust.

[10] This example is a slight variation on one considered by Wei Dai.

[11] Note that this problem is not ruled out by ditching the Vingean principle. Choosing the \$5 button is a consistent equilibrium; if you expect that's what you'll end up doing, then it is in fact what you end up doing.

Dai proposed a revised version of UDT to fix this problem, which he called UDT1.1 (making the original UDT retroactively UDT1, or as I prefer, UDT1.0). The suggested fix was to choose the whole policy at once, rather than choosing actions individually. This solves any potential coordination problems between instances. Unfortunately, it does not suit our purposes here, because it violates the Vingean principle: it requires the agent to plan everything at once, which is not realistic.[12] (In the rest of this paper, plain UDT refers to UDT1.0, but the reader is cautioned that this convention is not universal.)

The approach taken in this paper instead revolves around *communication*. Notice how UDT does not explicitly model agents with memory. If an agent does have memory, it needs to be modeled as part of the observation. This is a sort of communication between instances.

Example 2 (Memory problem)**.** [13] At time one, you will observe a red light or a green light. At time two, you will be offered an option to take one (or none) of two pills; one pill makes you say "red light" in response to any question for the duration of the problem, while the other does the same for "green light". After making this choice, you will get your memory wiped, and then (at time three) will be asked to report whether the light at time one was green or red, and rewarded for a correct answer.

Clearly, UDT lacks self-trust in this example; it will take the pill to modify its behavior. However, my contention is that this example is "unfair" in some sense: the pill was allowed to accomplish something which the agent's own memory was not allowed to do. In order to narrow things down to cases where self-trust can be treated as a rationality criterion, cases like this need to be ruled out. The intuition behind the present work is that lines of communication should "exactly parallel" lines of self-modification: if the agent has the ability to act like it remembers something with a self-modifying pill, then it should also be given the ability to remember normally. (All of this will be formalized later.)

Returning to Coordinated Buttons, the suggestion is this: before being copied, when considering whether to take the pill, you can can think to yourself "I should press the $10 button". Once you are copied, you and the copy can recall this thought and press the $10 button, secure in the knowledge that there would be no reason for your other instance to change its mind.

This allows communication to act as a release valve for pressures which would otherwise give rise to self-modification. Although it does imply some ability to predict other instances in some cases, it only does this to the extent that self-modification can be predicted; if a self-modifying action has effects which are uncertain in their particulars, then the corresponding self-communication will have similarly uncertain impacts. (This will become clearer when stated formally.) As such, I believe it respects the spirit of the Vingean principle.

Unfortunately, this idea will not be enough to carry us all the way.

Example 3 (Third Button problem)**.** [14] As in Coordinated Buttons, you are about to be copied, and you have the option of taking a pill which will cause you to press buttons labeled $10 for the duration of the problem. You also have the option of telling yourself to press $10 buttons. However, when you get into the red and green rooms, there will be a third button labeled $20. If both of you press $5, you both get $5; if both of you press $10, you both get $10; if one of you presses $10 and the other presses $20, you both get $20; in all other cases, you both receive $0. You have no way to randomize your choices, and you cannot tell yourself to do different things depending on the color of the room (your memory will be wiped in such a case).

This problem seems to be "fair" by the standards mentioned so far, yet UDT may still choose self-modification. Depending on its prior, UDT may still need to use the pill to choose $10. If it instead elects to tell itself to choose $10, then by the reasoning proposed earlier, each copy would trust that the other copy will follow this instruction; however, if that were true, *this would lead both copies to choose $20*, resulting in a payoff of $0.

As such, we will also need to rule out cases like this in order to achieve self-trust. To achieve this, we will need to deal with a conceptual flaw which arises in some interpretations of Yudkowsky's

[12] One might quibble over whether *deciding* everything at once violates my version of the Vingean principle, which only forbids requiring the agent to *predict* everything at once; however, it is clear that this should be forbidden for the same reason, namely that it is not cognitively realistic for an agent living in a large world.

[13] As far as I know, this example is novel, although it is simple enough that I would not be surprised to find an analogue in the literature.

[14] I believe this example is a novel contribution.

notion of fairness: naively, at least, self-modification is itself a decision, so Yudkowsky's notion of decision-determination fails to rule out universes which ask you to self-modify into an alphabetizer and punish you for not doing so (exactly the scenario Yudkowsky wants to rule out). We can resolve this problem by distinguishing between an agent's *external* observations and actions (these are the observations and actions most typically studied by decision theory), versus *internal* observations and actions (such as recalling or storing memories). A fair environment is, roughly, allowed to depend on external behaviors but not internal behaviors.

The remainder of the paper formalizes these ideas.

4 Mathematical Preliminaries

The mathematical formalism used here was significantly inspired by Finite Factored Sets [Gar21], although it differs considerably in the details, and does not attempt to deal with issues of time or causality.

Random variables will be modeled as partitions of a shared outcome set Ω. A partition X of Ω is a collection of nonempty, pairwise disjoint subsets of Ω whose union equals Ω. I will use uppercase letters to represent partitions (random variables) and matching lowercase letters for their parts (values).

For most of the random variables I consider, this approach captures the essential structure: each part $x \in X$ represents a possible value the variable can take, and each outcome $\omega \in \Omega$ belongs to exactly one part. It will be much more interesting to manipulate the partitions with standard partition operations, than it would be to follow the more standard approach and consider these random variables as real-valued functions of Ω.[15]

There is one exception to this: the UDT agent will have a utility function U which we need to treat as a real-valued function $U : \Omega \to \mathbb{R}$.

4.1 Refinement & Coarsening

The refinement relation provides a partial order on partitions:

Definition 1. For partitions X and Y, $X \leq Y$ ("X is a **refinement** of Y" or "Y is a **coarsening** of X") if and only if every $x \in X$ is a subset of some $y \in Y$.

Intuitively, X provides more information than Y when X refines Y, as X distinguishes information which Y groups together. I will also call Y a *subvariable* of X in this circumstance; X is like a larger system of which Y is a part. (EG, X could be the state of an entire motor and Y the state of an individual piston.)

Definition 2. If $X \leq Y$, the **projection** $[[\cdot]]_Y : X \to Y$ maps each $x \in X$ to the unique $y \in Y$ such that $x \subseteq y$. Projection of a function $f : Z \to X$ is defined as $[[f]]_Y(z) = [[f(z)]]_Y$.

Definition 3. For partitions X and Y:

- The **meet** $X \wedge Y$ is the unique coarsest common refinement of X and Y. It consists of all nonempty intersections $\{x \cap y : x \in X, y \in Y, x \cap y \neq \emptyset\}$.
- The **join** $X \vee Y$ is the unique finest common coarsening of X and Y. It has a part for every connected component of the relation $x \cap y \neq \emptyset$.

Subscripts like S_i will be used to denote members of a family of partitions $(S_i)_{i \in \mathcal{I}}$. The definitions of join and meet generalize naturally to act on families of partitions; this can be denoted $\bigvee_{i \in \mathcal{I}} S_i$ and $\bigwedge_{i \in \mathcal{I}} S_i$.

Definition 4. A function $f : X \to Y$ is a **restriction map** if $f(x) \subseteq x$ for all $x \in X$.

[15] A third approach would be to model random variables as σ-algebras, rather than as partitions. That choice is probably a bit superior mathematically, but the difference won't matter to me here.

4.2 Partition Factorization

In contrast to Finite Factored Sets, we will factor partitions rather than the underlying sets:

Definition 5. A partition X **factors as** (Y, Z) if and only if $X = \{y \cap z : y \in Y, z \in Z\}$. (Note that this implies $y \cap z$ is nonempty for all $y \in Y, z \in Z$.)

Equivalently, we can define factorization in terms of restriction maps: a partition X factors as Y, Z if and only if there exists a surjective restriction map $m : Y \times Z \to X$, namely, $m(y, z) = y \cap z$. (Note that this is also injective.)

$X = Y \wedge Z$ tells us that if we know Y and Z, we have enough information to compute X; however, this does not guarantee that arbitrary $y \in Y$ and $z \in Z$ are compatible. If they are incompatible, there will be no corresponding X. In contrast, "X factors as (Y, Z)" makes this further promise. That is: for every $y \in Y$, there exists a restriction map of type $Z \to X$, namely $m_y(z) = m(y, z)$; and similarly for every $z \in Z$ there exists a restriction map of type $Y \to X$, namely $m_z(y) = m(y, z)$.

This notion extends naturally to more than two factors, including families of partitions:[16]

Definition 6. X **factors as** $(X_i)_{i \in \mathcal{I}}$ if and only if there exists a surjective restriction map m from choice functions $c : (i : \mathcal{I}) \to X_i$ to X, namely $m(c) = \bigcap_{i \in \mathcal{I}} c(i)$.

Here and throughout, $(a : A) \to B_a$ denotes a dependent function type; the domain is A, and for every $a \in A$, $f(a) \in B_a$.

4.3 Functions

When a partition factors, it can be useful to compute functions into that partition in terms of functions into each factor:

Definition 7. A function $f : X \to Y$ **decomposes into** functions $f_i : X_i \to Y_i$ if:

1. Y factors as $(Y_i)_{i \in \mathcal{I}}$
2. Each X_i is a subvariable of X (not necessarily a factor)
3. $f(x) = \bigcap_{i \in \mathcal{I}} f_i([[x]]_{X_i})$ for all $x \in X$

I will denote the type of a partial function from X to Y as $X \rightharpoonup Y$. For $f : X \rightharpoonup Y$, $f(x) = \perp$ ("f is null at x") denotes the empty output (f takes no value for input x). $\mathrm{dom}(f)$ is the set of inputs receiving a non-null message.

5 Agents & Environments

In this section, we apply the above mathematical formalism to model agents. This approach takes significant inspiration from Critch's work on agent boundaries [Cri22] and Garrabrant's work on Cartesian Frames [GHLW21], though again, it differs considerably in the details. The restriction maps associated with a factorization play the role of the evaluation function which takes an agent and an environment and returns a world. However, I follow Critch in dividing the world into three parts rather than two: the agent's exterior, interior, and boundary. These are each partitions of Ω:

- I: The **interior** of the agent, representing persistent state (memories or more complex cognition).
- B: The **boundary** of the agent, implementing the decision procedure (handling input/output with the environment).
- E: The **external environment**, containing everything in the agent's exterior.

[16] This definition might look intimidating, but it is just a fancy way of saying that X factors into some larger number of factors $(X_1, X_2, ...)$.

For example, if I represents the SSD, each partition $i \in I$ is the set of worlds $\omega \in \Omega$ where the SSD holds a specific string of 1s and 0s.

As mentioned earlier, we assume the existence of a utility function $U : \Omega \to \mathbb{R}$. We will also assume a probability distribution $\mathbb{P}$ over Ω, which represents the subjective beliefs of our agent. We need to assume that the sigma-algebra for $\mathbb{P}$ includes all parts of every partition of interest. $\mathbb{E}(U|s)$ will represent the subjective expected utility of event $s \subseteq \Omega$. To ensure that expectations are well-defined, we can assume that utility is bounded $U(\omega) \in [0, 1]$ and stipulate that $\mathbb{E}(U|S) = -1$ when $\mathbb{P}(S) = 0$.[17]

5.1 Information Flow

Four variables mediate information flow between I, B, and E:

- $\dot{O}$: The **internal observation** represents memories or internal computations accessible to the decision procedure. $\dot{O}$ is the output of I, which flows into B (becoming one of the two inputs for B). $\dot{O}$ is a subvariable of both I and B.
- $\ddot{O}$: The **external observation** represents sense data. $\ddot{O}$ is the output of E, which flows into B (becoming the second input for B). $\ddot{O}$ is a subvariable of both B and E.
- $\dot{A}$: The **internal action** represents data to remember/process. $\dot{A}$ is the input to I, which flows out from B (one of the two outputs of B). $\dot{A}$ is a subvariable of both I and B. Together, $\dot{O}$ and $\dot{A}$ exhaust the information overlap between I and B: $I \vee B = \dot{O} \wedge \dot{A}$.
- $\ddot{A}$: The **external action**, representing motor commands. $\ddot{A}$ is the input to E, which flows out from B (the second output for B). $\ddot{A}$ is a subvariable of both B and E. $\ddot{O}$ and $\ddot{A}$ together exhaust the overlap between B and E; that is: $\ddot{A} \wedge \ddot{O} = B \vee E$.

We also define composite variables $O = \dot{O} \wedge \ddot{O}$ and $A = \dot{A} \wedge \ddot{A}$, representing the full observation (all inputs to B) and full action (all outputs of B) respectively. I will assume that O factors as $\dot{O}, \ddot{O}$ and A factors as $\dot{A}, \ddot{A}$. I will freely pun between O and $\dot{O} \times \ddot{O}$ and between A and $\dot{A} \times \ddot{A}$, so that for example $(\dot{o}, \ddot{o}) \in O$.

The status of these variables as "inputs" is justified by the existence of a "dynamic" capturing input-output relationships.

Assumption 1 (Internal Dynamic)**.** *I factors as $(\dot{A}, D_I)$ with restriction map $\iota : \dot{A} \times D_I \to I$.*

The variable D_I encodes how the interior processes inputs. Fixing a particular $d_I \in D_I$, we can compute $\iota_{d_I} : \dot{A} \to I$, which tells us how I responds to the internal action. Since $\dot{O}$ is a subvariable of I, this yields the input-output function of I, $[[\iota_{d_I}]]_{\dot{O}}$.

Assumption 2 (Environment Dynamic)**.** *E factors as $(\ddot{A}, D_E)$ with restriction map $\epsilon : \ddot{A} \times D_E \to E$.*

Similarly, D_E encodes how the environment processes external actions. Given $d_E \in D_E$, we get $\epsilon_{d_E} : \ddot{A} \to E$, determining the environment's response, including the external observations $\ddot{O}$ it will produce.

Assumption 3 (Boundary Dynamic)**.** *B factors as (O, D_B) with restriction map $\beta : O \times D_B \to B$.*

The dynamic D_B encodes the agent's decision rule. Given $d_B \in D_B$, we compute $\beta_{d_B} : O \to B$, which tells us the behavior of B as a function of its inputs.

Assumption 4 (Agent Dynamic Constraint)**.** *$I \wedge B$ factors as $(\ddot{O}, D_I \wedge D_B)$ with restriction map $\rho : \ddot{O} \times (D_I \wedge D_B) \to I \wedge B$.*

This constraint states that the combined agent state $I \wedge B$ is fully determined by the external observation and the joint dynamics of the interior and boundary. The joint dynamic will be denoted $D_{I,B} = D_I \wedge D_B$, with elements $d_{I,B}$, so that $\rho_{d_{I,B}} : \ddot{O} \to I \wedge B$ gives the joint state of I and B as a function of the external observation.

[17] Note that this stipulation ensures an agent will not choose a probability-zero action. This is undesirable, since it allows agents to avoid an action simply because they know they avoid that action, and for no other reason. An earlier draft of this paper had $\mathbb{E}(U|S) = 2$ in such cases instead; however, a reviewer pointed out that this causes problems with the "instances are believed to follow recommendations" assumption used in theorems 1 and 2.

To illustrate the advantage of modeling systems in this way, consider a system consisting of X and Y. These two parts of the system might have some complicated interaction. For example, X and Y might further factor into time-series variables $(X_t)_{t\in\mathbb{N}}$, $(Y_t)_{t\in\mathbb{N}}$, with causal relationships between them; then dynamics D_X and D_Y would represent the initial states and transition functions. If we were representing all of this in Bayesian networks, we would be forced to spell out such details. The current approach allows us to describe high-level functional relationships (containing apparent loops) without specifying all of that detail (which resolves the loops somehow).

Two more variables are important for understanding information flow, both subvariables of $\dot{O}$:

- $\hat{O}$: the **semantic observation**. This encodes all the information the agent is *supposed to* receive from the internal observation, when functioning normally.
- $\check{O}$: the **side channel**. This encodes aspects of the input which may modify the behavior of the agent, causing it to act "off-policy" IE causing its choices to contradict its decision rule.

For now, imagine that the variable $\check{O}$ has a "neutral" value that does nothing, and other values for specific modifications. This will be elaborated in more detail in Section 6.2.

Note that there is no assumption of factorization here; some combination of semantic observations and side-channel values may be incompatible, or inextricably linked. For example, 11001100 might be a semantically valid byte to read from memory, but one that triggers a hardware flaw in the CPU. There would be an $\hat{o} \in \hat{O}$ representing the intended semantics of 11001100, but it would be a subset of (or equal to) a non-neutral $\check{o} \in \check{O}$.

The assumption that $\check{O}$ is a subvariable of $\dot{O}$ is a significant one. It means that the environment cannot modify the agent forcibly. Furthermore, it means that the only way the agent can self-modify is through internal actions which nudge I to produce corrupting values of $\check{O}$. It could be interesting to also consider whether the agent will self-modify through the external environment if given opportunities to do so, but this seems antithetical to Yudkowsky's notion of decision-determination fairness: self-modification routes which pass through the external environment can be "viewed from outside" and can be incentivized or de-incentivized by the utility function. The main reason I am distinguishing between external and internal is to prevent this.

5.2 Notions of Policy

The usual notion of the 'policy' of an agent splits into several ideas in this setting.

Definition 8. A **full policy** is a function of type $O \to A$, mapping full observations to full actions.

Definition 9. The **effective policy** $\pi^\dagger \in \Pi^\dagger$ is the full policy actually implemented by the agent, namely the projection of the boundary dynamic to full actions, $[[\beta_{d_B}]]_A$.

I will freely pun between $\Pi^\dagger$ as a set of functions vs as a subvariable of D_B with a distinct value for each function, indicating which full policy the actual $d_B \in D_B$ implements.

I will sometimes use equations to denote the set of worlds satisfying the equation, EG, $\mathbb{P}(\Pi^\dagger(o) = a)$. Since uppercase letters represent variables and lowercase letters represent specific values, this is the probability that the true effective policy (whatever it may be) outputs specific action a on specific observation o.[18] When equations are comma-separated, EG $\mathbb{P}(e|\Pi^\dagger(o) = a, \Pi^\dagger(o') = a')$, this represents the intersection of the constraints.

Definition 10. The **chosen policy** $\pi^* \in \Pi^*$ is the full policy which *would be* implemented by the agent if the agent's decision procedure were never modified.

I assume that a subvariable representing the chosen policy exists, so that I can pun between Π^* as a set of functions vs as a variable, as for the effective policy. (We didn't need an extra assumption to do this for the effective policy, since its subvariable can be constructed.)

Definition 11 (External Policy)**.** The **external policy** $\ddot{\pi} : \ddot{O} \to \ddot{A}$ is what the agent's policy appears to be from an outside perspective, namely, $\ddot{\pi}(\ddot{o}) = [[\rho_{d_{I,B}}(\ddot{o})]]_{\ddot{A}}$.

[18] This is very similar to standard notation in statistics; EG, $\mathbb{P}(V = v)$ ordinarily denotes the probability that random variable V takes on the specific value v.

I will pun between $\ddot{\Pi}$ as the set of such functions or the subvariable of $D_{I,B}$ which chooses between those functions. (Again, this already exists without any need to assume it so, since it can be computed from existing variables.)

To understand the difference between the external policy and the full policy, imagine a robot with memory going back exactly 5 hours. The full policy treats actions as a function of the current external observations plus the memory. This allows us to ask what the robot would do if it remembered incorrectly. The external policy, on the other hand, views the robot's current action as a function of its entire history (a function which just so happens to depend only on the past five hours of that history, plus the current external observation). This only allows us to ask what would happen if the external observations were different.[19]

If we obey all the definitions and assumptions so far, I will say we have an *abstract decision structure*.

This gives us enough structure to define a version of Yudkowsy's decision-determination notion of fairness:

Definition 12 (Decision-Determination)**.** An abstract decision structure is **decision-determined** if and only if U depends only on E and E is conditionally independent of $D_{I,B}$ given $\ddot{\Pi}$:

$$\mathbb{P}(E = e | D_{I,B} = d_{I,B}) = \mathbb{P}(E = e | \ddot{\Pi} = [[d_{I,B}]]_{\ddot{\Pi}})$$

In words: the external policy is the only thing intrinsic to the agent that impacts the external environment, and the external environment is the only thing the agent cares about.

Note that this is not a philosophical claim that an agent *should* only care about the external environment. Rather, the point is that if an agent does have other preferences, it's an unfair problem from the perspective of impartially judging between decision procedures, since such preferences can introduce intrinsic biases towards one decision procedure and away from another.

6 Instances & Communication

So far, I have described I as a sort of internal state of the agent, which carries something like memories, or perhaps cognitive states which evolve over time in order to compute something. However, I have so far made it sound as if there is only a single decision being made, with a single external observation $\ddot{O}$ resulting in a single external action $\ddot{A}$. In order to model agents with multiple instances, I will factor the agent into *instances*. This resembles the notion of multiplicative subagents in Cartesian Frames.

6.1 Instance Structure

The agent could be factored in different ways. Here, I will treat *each external observation* as a different instance. For example, in the Coordinated Buttons problem in Section 3, the self in the red room is a different instance from the self in the green room *because the (external) observations are different* (red or green). If the observations of the two selves had been identical, then there would be no coordination problem for UDT. My choice to factor in terms of the external observation means that the red-room instance is 'the same instance' regardless of what they remember.

Definition 13 (Instance Factorization)**.** B factors as $(B_{\ddot{o}})_{\ddot{o} \in \ddot{O}}$, with restriction map $b : (\ddot{o} : \ddot{O}) \to B_{\ddot{o}} \to B$. Each $B_{\ddot{o}}$ represents the decision-making machinery that activates when $\ddot{o}$ is received. Correspondingly, B's subvariables factor:

- $\dot{A}$ factors as $(\dot{A}_{\ddot{o}})_{\ddot{o} \in \ddot{O}}$, with each $\dot{A}_{\ddot{o}}$ representing the internal action taken by instance $\ddot{o}$.
- $\ddot{A}$ factors as $(\ddot{A}_{\ddot{o}})_{\ddot{o} \in \ddot{O}}$, with each $\ddot{A}_{\ddot{o}}$ the external action of instance $\ddot{o}$.
- A factors as $(A_{\ddot{o}})_{\ddot{o} \in \ddot{O}}$. Each $A_{\ddot{o}}$ factors as $\dot{A}_{\ddot{o}}, \ddot{A}_{\ddot{o}}$.
- $\dot{O}$ factors as $(\dot{O}_{\ddot{o}})_{\ddot{o} \in \ddot{O}}$, each $\dot{O}_{\ddot{o}}$ being the internal observation of instance $\ddot{o}$.
- $\check{O}$ factors as $(\check{O}_{\ddot{o}})_{\ddot{o} \in \ddot{O}}$. Each $\check{O}_{\ddot{o}}$ is a subvariable of the corresponding $\dot{O}_{\ddot{o}}$.

[19] If the memory is fallible, this might show up as some uncertainty about the external policy, but would not uncertainty about the full policy.

- $\hat{O}$ factors as $(\hat{O}_{\ddot{o}})_{\ddot{o}\in\ddot{O}}$. Each $\hat{O}_{\ddot{o}}$ is a subvariable of the corresponding $\dot{O}_{\ddot{o}}$.
- Full policies π decompose into **instance policies** $\pi_{\ddot{o}} : \dot{O}_{\ddot{o}} \to \dot{A}_{\ddot{o}} \times \ddot{A}_{\ddot{o}}$ such that $\pi(\dot{o}, \ddot{o}) = \pi_{\ddot{o}}(\dot{o})$. Thus, each instance has its own policy mapping its observation to an appropriate action. Accordingly, Π^* factors as chosen instance policies $(\Pi^*_{\ddot{o}})_{\ddot{o}\in\ddot{O}}$, and $\Pi^\dagger$ factors as effective instance policies $(\Pi^\dagger_{\ddot{o}})_{\ddot{o}\in\ddot{O}}$.

For any variable S which factors as $(S_{\ddot{o}})_{\ddot{o}\in\ddot{O}}$, a subscripted value like $s_{\ddot{o}}$ denotes a value of the factor $S_{\ddot{o}}$ (so $s_{\ddot{o}} \in S_{\ddot{o}}$).

This way of factoring an agent into instances makes sense for UDT, and for any functional decision theory (FDT) [YS18], because the philosophy behind these decision theories identifies the "choice" as the abstract mathematical decision: if an exact physical copy of me is placed in a physically identical room, there is only one "choice" being made. This notion of "instance" fits less well with the philosophy behind other decision theories. From this point on in the paper, the mathematical framework is specific to the theorems I want to prove about UDT, rather than providing a general formalism for evaluating decision theories.

6.2 Internal Communication Structure

The factorization of internal messages allows $\dot{O}$ to act as a "bundle of messages" with one message for each instance. I will assume a very specific message semantics:

Definition 14 (Message Semantics)**.** Each $\hat{o}_{\ddot{o}} \in \hat{O}_{\ddot{o}}$ can be understood as either recommending a specific external action to instance $\ddot{o}$, or remaining silent. The partial function $s_{\ddot{o}} : \hat{O}_{\ddot{O}} \rightharpoonup \ddot{A}$ interprets the semantics for us. This gives rise to a second partial function of interest, $r_{\hat{o}} : \ddot{O} \rightharpoonup \ddot{A}$, the **recommendation**; $r_{\hat{o}}(\ddot{o}) = s_{\ddot{o}}(\hat{o}_{\ddot{o}})$. The variable R represents the possible recommended external policies; it is a subvariable of $\dot{O}$.

I will pun between understanding R as a variable and as a set of functions. I will write $\ddot{\Pi} = R$ to indicate that the recommendations are followed, IE, whatever their specific values turn out to be, $\forall \hat{o} \in \hat{O}, \forall \ddot{o} \in \mathrm{dom}(r_{\hat{o}}) : \ddot{\pi}(\ddot{o}) = r(\ddot{o})$. The notation $[\ddot{\Pi} = R]_{-\ddot{o}}$ indicates that all of the recommendations are being followed except perhaps at instance $\ddot{o}$.

Definition 15 (Side-Channel Impact)**.** Each $\check{o}_{\ddot{o}} \in \check{O}_{\ddot{o}}$ either forces the receiving instance to take a specific external action, or does nothing. The partial function $p_{\ddot{o}} : \check{O}_{\ddot{o}} \rightharpoonup \ddot{A}$ indicates what action is forced by a specific value of $\check{O}_{\ddot{o}}$. This gives rise to $q_{\check{o}} : \ddot{O} \rightharpoonup \ddot{A}$, the **forced external policy**: $q_{\check{o}}(\ddot{o}) = p_{\ddot{o}}(\hat{o}_{\ddot{o}})$. The variable P represents the possible forced external policies; it is a subvariable of $\dot{O}$.

Definition 16 (Modification)**.** If $\mathrm{dom}(q_{\check{o}}) \neq \emptyset$, a **modification** has occurred. The **modification probability** of an event e is $m(e) = \mathbb{P}(\mathrm{dom}(q_{\check{o}}) \neq \emptyset | e)$.

I will call an abstract decision structure which satisfies all of the conditions in this section a *concrete decision structure*.

This formalism for self-modification has several advantages over the formalism used in [DHR25]. In that paper, self-modifications were always achieved by single actions whose impact on the effective policy was deterministic. Here, there is merely some probabilistic relationship between internal actions and (potentially self-modifying) internal observations. Self-modification can require multiple actions by different instances, and can have uncertain effects. Finally, although the framework does forbid the environment from modifying the agent, it can somewhat distinguish between self-modification vs mere modification; it may be that all actions available have nonzero modification probability (there is an unavoidable chance of modification).

7 Avoiding Self-Modification

Consider an agent with chosen instance policy π^* determined by UDT:

$$\pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = \arg\max_{a_{\ddot{o}} \in A_{\ddot{o}}} \mathbb{E}[U | \Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = a_{\ddot{o}}]$$

This decision procedure seems particularly well-suited to decision-determined problems; indeed, Dai formulated it in response to Yudkowsky's ideas [Dai09]. Therefore, we seek to prove a version of self-trust for UDT.

Definition 17. An action $a_{\ddot{o}}$ is **minimally modifying** at $\dot{o}_{\ddot{o}}$ if and only if $m(\Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = a_{\ddot{o}}) = \min_{a'_{\ddot{o}} \in A_{\ddot{o}}} m(\Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = a'_{\ddot{o}})$.

We wish to show that, for decision-determined concrete decision structures, non-minimally-modifying actions are never strictly preferred by UDT. However, the Coordinated Buttons problem discussed in Section 3 provides a counterexample to such a theorem. As suggested in that section, we need to further assume that the agent has the option of communicating to itself rather than self-modifying.

Definition 18 (Communicative Alternative)**.** For any instance-action $a_{\ddot{o}} \in A_{\ddot{o}}$ and internal observation $\dot{o}_{\ddot{o}} \in \dot{O}_{\ddot{o}}$, a **communicative alternative** $ca_{\dot{o}_{\ddot{o}}}(a_{\ddot{o}}) \in A_{\ddot{o}}$ satisfies:

1. $ca(a_{\ddot{o}})$ is minimally modifying at $\dot{o}_{\ddot{o}}$
2. For all external policies $\ddot{\pi}$:

$$\mathbb{P}(\ddot{\pi} \mid \Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = a_{\ddot{o}}) = \mathbb{P}(\ddot{\pi} \mid \Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = ca(a_{\ddot{o}}), \ddot{\Pi} \equiv R)$$

We can prove that this action is just as good:

Lemma 1 (Communicative Expectation)**.** *If a concrete decision structure is (1) decision-determined and (2) has communicative alternatives, then*

$$\mathbb{E}\big(U|\Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = a_{\ddot{o}}\big) = \mathbb{E}\big(U|\Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = ca(a_{\ddot{o}})\big)$$

Proof: By (1), we can decompose the expectations as follows:

$$\mathbb{E}\big(U \mid \Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = a_{\ddot{o}}\big) = \sum_{\ddot{\pi} \in \ddot{\Pi}} \mathbb{E}\big(U \mid \ddot{\pi}\big)\mathbb{P}\big(\ddot{\pi} \mid \Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = a_{\ddot{o}}\big)$$

$$\mathbb{E}\big(U \mid \Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = ca_{\dot{o}_{\ddot{o}}}(a_{\ddot{o}}), \ddot{\Pi} = R\big) = \sum_{\ddot{\pi} \in \ddot{\Pi}} \mathbb{E}\big(U \mid \ddot{\pi}\big)\mathbb{P}\big(\ddot{\pi} \mid \Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = ca_{\dot{o}_{\ddot{o}}}(a_{\ddot{o}}), \ddot{\Pi} = R\big)$$

Since the distributions match by (2), the expected utilities are equal. □

We can now prove our main theorem:

Theorem 1 (Self-Trust)**.** *If a concrete decision structure is (1) decision-determined, (2) has communicative alternatives, and (3) has $\mathbb{P}(\ddot{\Pi} = R) = 1$, then non-minimally-modifying actions can never be strictly preferred by UDT. That is:*

$$m(\Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = a_{\ddot{o}}) > \min_{a'_{\ddot{o}} \in A_{\ddot{o}}} m(\Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = a'_{\ddot{o}})$$

$$\Longrightarrow$$

$$\mathbb{E}\big(U|\Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = a_{\ddot{o}}\big) \leq \max_{a'_{\ddot{o}} \in A_{\ddot{o}}} \mathbb{E}\big(U|\Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = a'_{\ddot{o}}\big)$$

Proof: Let $\dot{o}_{\ddot{o}} \in \dot{O}_{\ddot{o}}$, $\ddot{o} \in \ddot{O}$, and $a_{\ddot{o}} \in A_{\ddot{o}}$ be such that:

$$m(\Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = a_{\ddot{o}}) > \min_{a'_{\ddot{o}} \in A_{\ddot{o}}} m(\Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = a'_{\ddot{o}})$$

By (2) and the lemma,

$$\mathbb{E}\big(U|\Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = a_{\ddot{o}}\big) \leq \mathbb{E}\big(U|\Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = ca_{\dot{o}_{\ddot{o}}}(a_{\ddot{o}}), \ddot{\Pi} = R\big)$$

By (3), this proves the desired result. □

At this point, I expect many readers will accuse me of foul play. Condition (3) may strike you as assuming too much. Surely we can't just assume that instances will follow recommendations? Remember, though, $\mathbb{P}$ is the subjective probability distribution of the agent. I am not assuming that

instances follow recommendations; rather, I am assuming *the agent expects* its instances to follow recommendations.[20]

The intuition behind this theorem, for me, is that we can't get coordination from nothing.[21] Instead, we have to start by assuming coordination in a smaller domain, namely language. This modest coordination is leveraged to detangle other coordination problems.

Clearly, in order for language to do its job, we have to make *some* assumption that the language is mutually intelligible. Since this particular language is imperative, how else can we assume intelligibility, other than assuming at least some expectation that recommendations will be followed?

Still, is this too much? Some recommendations are bad. Believing that all recommendations are followed may not even be consistent with understanding one's instances as rational. The next section discusses conditions under which this assumption is reasonable.

8 Following Advice

We want to use messages as a "release valve" for self-modification: if all instances of an agent share the same goals, intuitively, the agent should be able to come up with a plan and remember that plan rather than forcing the plan on itself.

However, the Third Button problem discussed in Section 3 shows that this doesn't always work. In some cases, everyone trusting others to follow the plan leads to "defectors" who spoil the plan, even in cases where everyone shares the same goals and beliefs!

Clearly, we need some additional assumptions to rule out such counterexamples.

Definition 19 (Stability)**.** Write $[\Pi^* = r_{\hat{o}}]_{-\ddot{o}}$ to mean that the *particular* recommendation $r_{\hat{o}}$ is being followed (in contrast to $[\ddot{\Pi} = R]_{-\ddot{o}}$, which means the recommendation is being followed, whatever it is). A recommendation $r_{\hat{o}}$ is **stable** if and only if, for all external observations $\ddot{o} \in \mathrm{dom}(r_{\hat{o}})$, all internal actions $\dot{a}_{\ddot{o}} \in \dot{A}_{\ddot{o}}$, and all $\ddot{a}'_{\ddot{o}} \neq r_{\hat{o}}(\ddot{o})$:

$$\mathbb{E}\Big(U|\Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = \big(\dot{a}_{\ddot{o}}, r_{\hat{o}}(\ddot{o})\big), [\ddot{\Pi} = r_{\hat{o}}]_{-\ddot{o}}\Big) > \mathbb{E}\big(U|\Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = (\dot{a}_{\ddot{o}}, \ddot{a}'_{\ddot{o}}), [\ddot{\Pi} = r_{\hat{o}}]_{-\ddot{o}}\big)$$

That is, conditioned on all other advice being followed, following the advice you receive is better than ignoring it.

Definition 20 (Internally-Driven Recommendations)**.** Recommendations are **internally driven** if they cannot be manipulated by selection of different external actions:

$$\forall \dot{o} \in \dot{O}, \ddot{o} \in \ddot{O}, \dot{a} \in \dot{A}, \ddot{a}, \ddot{a}' \in \ddot{A} : \mathbb{P}\big(r_{\dot{o}}|\Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = (\dot{a}_{\ddot{o}}, \ddot{a}_{\ddot{o}})\big) = \mathbb{P}\big(r_{\dot{o}}|\Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = (\dot{a}_{\ddot{o}}, \ddot{a}'_{\ddot{o}})\big)$$

This assumption frees instances to optimize internal actions and external actions separately.

Theorem 2 (Advice-Following)**.** *If a concrete decision structure (1) has $\mathbb{P}(\ddot{\Pi} = R) = 1$, (2) has internally-driven recommendations, (3) has recommendations which are stable with probability one, then for (4) $s_{\ddot{o}}([[\dot{o}]]_{\hat{O}}) \neq \bot$, and for all $\ddot{a}_{\ddot{o}} \neq s_{\ddot{o}}([[\dot{o}]]_{\hat{O}})$, we have*

$$\mathbb{E}\Big(U\Big|\Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = \big(\dot{a}_{\ddot{o}}, s_{\ddot{o}}([[\dot{o}]]_{\hat{O}})\big)\Big) > \mathbb{E}\big(U\big|\Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = (\dot{a}_{\ddot{o}}, \ddot{a}_{\ddot{o}})\big)$$

That is, UDT will strictly prefer to follow advice. If we further assume UDT is the agent's decision rule, and $s_{\ddot{o}}([[\dot{o}]]_{\hat{O}}) \neq \bot \implies p_{\ddot{o}}([[\dot{o}]]_{\hat{O}}) = \bot$ (instances never receive recommendations and modifications at the same time), then $\ddot{\Pi} = R$ in fact.

Proof: We wish to show that $\arg\max_{\ddot{a}_{\ddot{o}} \in \ddot{A}_{\ddot{o}}} \max_{\dot{a}_{\ddot{o}} \in \dot{A}_{\ddot{o}}} \mathbb{E}\big(U \mid \Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = (\dot{a}_{\ddot{o}}, \ddot{a}_{\ddot{o}})\big) = r_{[[\dot{o}]]_{\hat{O}}}(\ddot{o})$. Writing $r_{\hat{o}} - \ddot{o}$ for the intersection of all $\hat{o}_{\ddot{o}'}$ except $\ddot{o}' = \ddot{o}$, the left-hand side can be decomposed as

$$\arg\max_{\ddot{a}_{\ddot{o}} \in \ddot{A}_{\ddot{o}}} \max_{\dot{a}_{\ddot{o}} \in \dot{A}_{\ddot{o}}} \sum_{r_{\hat{o}'} - \ddot{o} \text{ for } \hat{o}' \in \hat{O}} \mathbb{E}\big(U\big|\Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = (\dot{a}_{\ddot{o}}, \ddot{a}_{\ddot{o}}), r_{\hat{o}'} - \ddot{o}\big)\mathbb{P}\big(r_{\hat{o}'} - \ddot{o}\big|\Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = (\dot{a}_{\ddot{o}}, \ddot{a}_{\ddot{o}})\big)$$

[20] Also note that since the contents of r are uncertain, this does not violate the Vingean Principle, as it does not imply that instances know specifically what each other will do. In particular, it does not imply that any instance knows what it itself will do, even though that instance can see its own message, since decisions are made updatelessly.

[21] At least, it is too expensive; we can't afford a big search to find the optimal policy, as proposed by UDT1.1.

Let $\dot{a}_{\ddot{o}}$ denote the internal instance action which is best given the best external instance action:

$$= \arg\max_{\ddot{a}_{\ddot{o}} \in \ddot{A}_{\ddot{o}}} \sum_{r_{\hat{o}'} - \ddot{o} \text{ for } \hat{o}' \in \hat{O}} \mathbb{E}\big(U \big| \Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = (\dot{a}_{\ddot{o}}, \ddot{a}_{\ddot{o}}), r_{\hat{o}'} - \ddot{o}\big) \mathbb{P}\big(r_{\hat{o}'} - \ddot{o} \big| \Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = (\dot{a}_{\ddot{o}}, \ddot{a}_{\ddot{o}})\big)$$

By (1),

$$= \arg\max_{\ddot{a}_{\ddot{o}} \in \ddot{A}_{\ddot{o}}} \sum_{r_{\hat{o}'} - \ddot{o} \text{ for } \hat{o}' \in \hat{O}} \mathbb{E}\big(U \big| \Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = (\dot{a}_{\ddot{o}}, \ddot{a}_{\ddot{o}}), [\ddot{\Pi} = r_{\hat{o}}]_{-\ddot{o}}\big) \mathbb{P}\big(r_{\hat{o}'} - \ddot{o} \big| \Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = (\dot{a}_{\ddot{o}}, \ddot{a}_{\ddot{o}})\big)$$

By (2), $\mathbb{P}\big(r_{\hat{o}'} - \ddot{o} \big| \Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = (\dot{a}_{\ddot{o}}, \ddot{a}_{\ddot{o}})\big)$ is constant in $\ddot{a}_{\ddot{o}}$, so

$$= \arg\max_{\ddot{a}_{\ddot{o}} \in \ddot{A}_{\ddot{o}}} \sum_{r_{\hat{o}'} - \ddot{o} \text{ for } \hat{o}' \in \hat{O}} \mathbb{E}\big(U \big| \Pi^*(\dot{o}_{\ddot{o}}, \ddot{o}) = (\dot{a}_{\ddot{o}}, \ddot{a}_{\ddot{o}}), [\ddot{\Pi} = r_{\hat{o}}]_{-\ddot{o}}\big)$$

By (3) and (4),

$$= r_{[[\dot{o}]]_{\hat{O}}}(\ddot{o})$$

Furthermore, if the agent's decision procedure is in fact UDT, and if instances never receive recommendations and modifications at the same time (which could override UDT's preference), then the agent will in fact follow recommendations. □

What this theorem shows is that (under some additional conditions) *if the agent believes it will follow its own advice, then it in fact will do so.* This is similar to showing that there exist equilibria where advice is followed. The reason I did not use an equilibrium analysis is that my interpretation of the Vingean principle outlaws that sort of reasoning.

Theorem 2 helps us understand when $\mathbb{P}(\ddot{\Pi} = R) = 1$ is a reasonable thing for the agent to believe. Although $\mathbb{P}$ is subjective, and the Vingean principle invites us to consider bounded-cognition agents for whom $\mathbb{P}$ cannot perfectly reflect reality, still, beliefs should reflect reality to some extent. Without a result similar to Theorem 2, Theorem 1 is dissatisfying because it predicates self-trust on the belief that agents expect to follow their own advice, but provides no reason to think it is plausible (or even consistent) for agents to believe this. Theorem 2 shows that this belief (under some further conditions) is a self-fulfilling one.

9 Conclusion

Yudkowsky's notion of fairness (decision-determination) arguably had a conceptual flaw. Yudkowsky sought to specify the condition under which self-trust can be viewed as a rationality condition, used to rule out decision procedures which would predictably seek to modify themselves. He proposed decision-determination as that condition. However, self-modification is inherently a decision itself. This means decision-determination does not rule out decision problems which directly incentivize self-modification.

I interpreted decision-determination in a way which avoids this problem, by making an explicit distinction between internal actions and external actions. This allowed me to provide a formal definition of decision-determination, codifying the idea of a "fair problem" as one where the universe doesn't care about your internal makeup (ie, how the decisions get made) except in so far as it impacts your pattern of choices.

However, I found that (under my interpretation), decision-determination still isn't enough. Section 3 provided three counterexamples: cases where self-trust can fail without it appearing to be the fault of the decision procedure (hence, apparently "unfair" decision-determined problems).

Theorem 1 showed self-trust for UDT under three conditions: decision-determination, the existence of communicative alternatives, and belief that one will take one's own advice. Theorem 2 clarified the plausibility of that third condition, by showing that belief-in-advice-following implies advice-following in fact, under the further three conditions of internally-driven recommendations, stable recommendations, and recommendations being mutually exclusive with self-modifications.

It is not clear to me which of these assumptions, exactly, should be fairness conditions. I would suggest that fairness consists of at least decision-determination, communicative alternatives, and stable recommendations. When those are not present, failures of self-trust really do not feel like the agent's fault to me.

I would further suggest that something similar to $\mathbb{P}(\ddot{\Pi} = R) = 1$ seems necessary. This provides a sort of "self-esteem" needed to keep an agent going. If the agent does not believe that it will follow through on its plans, I think a rational decision procedure cannot save it.[22] Failure of self-trust in these cases may be the agent's fault, but it is not the decision procedure's fault.

This question has some philosophical weight to it. As I mentioned earlier, the assumption $\mathbb{P}(\ddot{\Pi} = R) = 1$ is an attempt to formalize the intelligibility of the shared language. Elevating this to the status of a fairness condition would say something about rationality requiring a cultural context of shared norms. A generalized theorem which deals with a richer language (EG, declarative sentences in addition to imperative ones, and more complex sentences of both sorts) could shed further light on this.

It is not hard to see how mutual intelligibility relates to AI safety, although the theorems presented here do not yet provide useful advice in that regard.

It would be very interesting to try to transform the ideas in this paper into a representation theorem, providing general conditions under which we *can* model agents this way, rather than merely constructing things by fiat.

I am eager for this work to be adapted to a setting which models computational uncertainty properly, as discussed in [DHR25]. The theorems here and in that paper visibly suffer from the attempt to adhere to the Vingean principle without a real model of bounded cognition.

Another (possibly related) direction would be to generalize these theorems to individuals who do not share identical goals and beliefs. Even better would be to generalize to the case where the sigma-algebra is not shared (no shared ontology). Ultimately, we want models which can say useful things about trust in a collection of agents including AIs and humans.

References

[Cri22] Andrew Critch. ≪boundaries≫ sequence (index post). https://www.lesswrong.com/s/LWJsgNYE8wzv49yEc, 2022.

[Dai09] Wei Dai. Towards a new decision theory. https://www.lesswrong.com/posts/de3xjFaACCAk6imzv/towards-a-new-decision-theory, 2009.

[Dai10] Wei Dai. Explicit optimization of global strategy (fixing a bug in UDT1). https://www.lesswrong.com/posts/g8xh9R7RaNitKtkaa/explicit-optimization-of-global-strategy-fixing-a-bug-in, 2010.

[DHR25] Abram Demski, Norman Hsia, and Paul Rapoport. Understanding trust. In *Proceedings of ILIAD*, Berkeley, California, 2025. Presented at ILIAD Conference, August 28–September 3, 2024.

[FLO02] Shane Frederick, George Loewenstein, and Ted O'Donoghue. Time discounting and time preference: A critical review. *Journal of Economic Literature*, 40(2):351–401, 2002.

[Gar21] Scott Garrabrant. Temporal inference with finite factored sets. *arXiv preprint arXiv:2109.11513*, 2021.

[GHLW21] Scott Garrabrant, Daniel A. Herrmann, and Josiah Lopez-Wild. Cartesian frames. *arXiv preprint arXiv:2109.10996*, 2021.

[Gus22] Johan E. Gustafsson. *Money-Pump Arguments*. Cambridge Elements in Decision Theory and Philosophy. Cambridge University Press, 2022.

[Mac89] Mark J. Machina. Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature*, 27(4):1622–1668, 1989.

[PCKng] Richard Pettigrew, Catrin Campbell-Moore, and Jason Konek. Choosing how to choose. *Theory and Decision*, forthcoming.

[22] Of course an *irrational* decision procedure can save it; you can take actions consistent with the optimal plan irrationally. This decision procedure will perform worse in subjective expectation, but better on average in any case where it is consistently implemented.

[Pol68] Robert A. Pollak. Consistent planning. *The Review of Economic Studies*, 35(2):201–208, April 1968.

[Str55] Robert H. Strotz. Myopia and inconsistency in dynamic utility maximization. *The Review of Economic Studies*, 23(3):165–180, 1955.

[Vin22] Susan Vineberg. Dutch Book Arguments. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition, 2022.

[YH13] Eliezer Yudkowsky and Marcello Herreshoff. Tiling agents for self-modifying ai, and the löbian obstacle. *MIRI technical report*, 2013.

[YS18] Eliezer Yudkowsky and Nate Soares. Functional decision theory: A new theory of instrumental rationality. Technical report, Machine Intelligence Research Institute, 2018.

[Yud10] Eliezer Yudkowsky. Timeless decision theory. *The Singularity Institute technical report*, 2010.