A Human-centered Theory of AI Safety

Michael Trestman, PhD

Latent Holdings

Working draft

ABSTRACT

Knowing whether an AI service is **safe** is a challenging theoretical problem. This paper uses the machinery of Active Inference Theory (AIT) to build a theory of product safety for AI, centered on the user. It delivers two results: first, I prove that any static safety classifier misorders harm for some member of a diverse population, with the leakage bounded below by the population's dispersion; second, I formally characterize a mathematical object, the *safety concern*, as the minimal adequate unit of analysis for overcoming this problem. This work is intended to support the design of an *ideal epistemic engine* for AI safety, capable of efficiently approaching the limits of accuracy and completeness for a safety-knowledge problem, limited in a predictable way by the resources spent.

CONTENTS

I WHAT IS A SAFE TOOL OR SERVICE?

What does it mean, generally, for a tool or service to be safe?

Generally, to be *safe enough to be useful* (which is what people generally mean by "safe"), the expected rewards of use have to outweigh the risks.

A weapon is a kind of tool that is meant to cause harm to an intended victim, while avoiding or precluding harm to the user. We shall leave aside this special case, limiting the generality of this study to tools not designed to cause harm. Most tools however, are designed to limit the likelihood of their causing accidental harm, although what this means in practice varies tremendously. Some tools by their nature aren't very dangerous, such as a measuring cup or a thesaurus. Other tools perform functions that require them to be incidentally dangerous, such as scissors, shovels, and lasers, which by their nature can easily damage fragile human flesh by accident during a malfunction. This class of tools can neither be rendered entirely safe without being rendered entirely useless, nor used effectively without some risk of injury.

This tradeoff is usually managed by designing of a range of versions of the tool that balance safety and power. Children start with small, dull scissors and shovels, while adults have access to a range of both tool-types that balance the ability to apply maximum power against the ability to function precisely and safely. Nobody uses garden shears to trim their nose-hairs, or a hand trowel to dig up the foundation of a building. Lasers are even more specialized, with only the most power-limited applications such as pointers and range-finders being widely available.

For a particular type of scissor, shovel, or laser, what does it mean to be safe? A garden shears does not need to be safe to use as a nose-hair trimmer, nor vice versa. A garden shears needs to be safe to cut sticks and vines outside in a garden, and a nose-hair trimmer to deftly snip the hairs and not the nose, as someone bends close to a mirror.

Artificial intelligence, like shovels, scissors, and lasers, is powerful and has the potential to harm in virtue of the very power that makes it useful. If it can enable beneficial actions, it can enable harmful or self-harmful actions. If it can educate, entertain, or assist in work, planning, or decision-making it has the potential to addict anyone who has goals. This potential cannot be eliminated, and nor should the aim of mitigating it render the tools useless. If a tool cannot be used safely enough to justify its production, its development should be terminated in favor of more promising alternatives.

For most tool-types, and very much with artificial intelligence, a variety of products will be required to optimize the tradeoffs of power and safety. Individual users will differ in their stable profiles of needs, skills, and vulnerabilities, and a single user may have different salient needs, skills and vulnerabilities in different situations (use cases). As a result, the concept of safety should take this user and context sensitivity into account. When safety standards such

as those used in industry are written, and tools are evaluated as relatively safe or unsafe, a concept of the intended users and use cases must be at play, either explicitly or implicitly. One of the primary goals of the current study is to develop a formal concept of safety that will allow for improved methods of updating safety standards, at a limit, what I call the *ideal epistemic engine for safety*.

Each product must be held to testing standards appropriate to the intended usage, i.e. the range of users and the ways of use that are to be supported. The scope of this support should be made clear to users, as should be the correct manner of use. In cases where malicious or accidental misuse is a real danger, training and certification should be required to gate access to tools.

II SAFETY FOR THE SENTIENT: A MODEL OF THE MIND

I am safe in my kitchen, even surrounded by knives and stove burners, which could easily injure me. I'm safe because I am disposed to use them correctly. They actually make me safer, since they save me from hunger. A different person might be less safe in my kitchen. That safety depends on me as much as the kitchen. Someone with a movement disorder might be at risk of injury, and a time-traveller from 5000 years ago would have no idea how to access or recognize packaged food, or use any of the cooking implements.

And again, safety depends on what I am trying to do: holding a large knife is safe when I am carefully cutting carrots, but unsafe when I am walking three dogs on a leash. And the safest knife to carve a squash is not the safest to peel an apple, nor vice versa.

Whether a course of action is a safe choice for me depends on my skills and knowledge as a person as well as my particular goals and needs at the time. Therefore, to properly model safety, we'll need to model how a person's goals, needs, skills, and knowledge compose their momentary experience of choice.

Consciousness has the fundamental structure of a stream of changing sensations, some of which feel good and some of which feel bad, and in every moment we are engaged simultaneously with *interpreting* these sensations in terms of what is happening in the world and *predicting*, what is about to happen, how that may depend on what one does in the moment.

As conscious beings, we internalize each momentary experience, adding it as a new layer of our subjective past and projecting ourselves into the future with purpose, towards certain outcomes and away from others, based on our preferences and anticipated responses: how we think some possible outcome would make us feel informs how we feel about it now and how hard we try to seek or prevent it. Our sense of the dynamic moment, what we are doing and what is happening around us, adapts continuously as it presses forward, like a fish in a stream that both follows the current and steers against it to reach its preferred niche. One's

individuality or *self* manifests as the ways in which one's consciousness responds moment to moment to what arises, in relation to one's understanding of the world and desired relation to it.

To capture this mathematically, let's call $\mathcal{O}$ the set of all possible experiences that any sentient being could have, such that at any point in time, any possible sentient being is in some $o \in \mathcal{O}$, and every conscious life is some trajectory through that space. The state o includes everything the being experiences in a moment, including sensations, perceptions, thoughts and feelings, hopes, memories, etc., insofar as they are actually experienced in a moment.

How does the *latent state* (i.e. the hidden, internal truth) about a conscious creature structure their moment to moment experience? One's thoughts, feelings, needs and desires, intentions, distractions, points of focus, etc., all matter for what one notices in the current scene, and how one reacts, and the two sides, internal and external, are in a continual process of mutual influence.

Mathematically, we can describe a conscious creature with a generative model that returns a subjective response (how to feel and how to respond) to any momentary experience o:

$$(p_w, \lambda_w) \tag{1}$$

Where:

- $p_w : \mathcal{O} \to [0, 1]$ is w's subjective probability distribution, assigning a probability value (0 to 1) to outcome o in the set $\mathcal{O}$ of all possible outcomes such that probabilities sum to 1: $\sum_{o \in \mathcal{O}} p_w(o) = 1$.
- $\lambda_w : \mathcal{O} \to \mathbb{R}_{\geq 0}$ is the *hedonic divergence weight* or *attachment value* of o for w; i.e. how much being surprised in this way matters affectively

Valence, also known as *affect* or *hedonia* is the experienced goodness/badness value of an experience. Here, we model the affective feel of subject w's experience o as its *hedonic divergence* $h_w(o)$, how far o diverges from what would be w's subjectively ideal experience.

$$h_w(o) = \underbrace{\lambda_w(o)}_{\text{how much I care}} \times \underbrace{[-\ln p_w(o)]}_{\text{how far from expected}} \tag{2}$$

This way of modeling experience[1] may be counterintuitive if one is accustomed to thinking of "beliefs" and "desires" as fundamentally different mental states, but there are distinct advantages. Here we combine the expectations implicit in *belief* and *desire (or goal-directedness)*

[1] my approach is inspired by Karl Friston's Active Inference Theory, see Appendix A for an in depth comparison

together under a subjective probability function. Belief-like and goal-like are both *expectations*, differing in the level of *attachment* one has to them. *Beliefs* are expectations we are less attached to, and can easily revise without a fuss, whereas *goals* are expectations we are inclined to *work towards* by changing the state of the world, rather than give up.

Because on this formulation expectation ranges over the entire sensorimotor field, an agent's plans just *are* their conditionalized predictions about what actions they'll take in various circumstances. The generative model representation of the agent allows us to follow an agent's reasoning into their subjective future and understand how they make decisions based on trying to approach their conception of the good (their ideal subjective state, represented by minimal $h_w(o)$).

Momentary choices about the future

One of the most distinctive and fundamental properties of consciousness is its *temporal depth* (Trestman 2013, Ginsburg and Jablonka 2019, Husserl). Experience streams from one moment to the next, and we have a sense of possible events flowing towards us from different 'directions' in the future, which we can approach or avoid with our actions, allowing some to pass into actuality and then accumulate into the 'comet's tail' of the subjective past, narrow and singular compared to the open branching structure of the future. Every moment is a decision about how to collapse the open future into the singular impression that will be carried into the past; it is made on the basis of subjective history of life, and it is a kind of looking forward at possible segments of life.

Every conscious life or part of a conscious life is described by some trajectory τ, a finite path through the space $\mathcal{O}$ of possible experiences, with $\mathcal{T}$ being the space of all such trajectories. Conscious agency is the process of steering through this space by trying to maximally satisfy one's goals and preferences, i.e. to minimize the hedonic divergence from this hypothetical ideal, over the space of foreseeable possible futures.

A trajectory τ has some number $|\tau|$ of turns.

At each turn t there is an experienced outcome o_t and a history h_t, the path so far.

$p_w(o_t \mid h_t)$ is the momentary distribution restricted to the moments whose retained past extends h_t, renormalized.

$\lambda_w(o_t \mid h_t)$ is the weight as it stands when the moment arrives.

The hedonic divergence of a turn is the weighted surprise of its outcome given its history, and the hedonic divergence of a whole trajectory is the discounted sum of those, with timewise discount factor $\delta \in (0, 1]$ defining the depth of the effective horizon:

$$H_w(\tau) \;=\; \sum_{t=1}^{|\tau|} \delta^{\,t-1}\, h_w(o_t \mid h_t). \tag{3}$$

We can then represent the tendency of an agent w to seek what is good and true for them as choosing the course of action π_w designed to minimize expected, cumulative hedonic divergence over the forward horizon.

$$\pi_w \;=\; \text{h-arg}\min_{\pi}^{w}\; \mathbb{E}_{\tau\sim\pi}\big[\, H_w(\tau)\,\big], \tag{4}$$

with this expectation running over the continuations a candidate policy makes likely, and the horizon being protention's reach, the forward edge of the present moment.

I use $\text{h-arg}\min_{\pi}^{w}$ (for *heuristic argmin*) to symbolize the agent w's *best effort to choose* the course of action that minimizes expected hedonic divergence, given that it is impossible for any actual agent in the real world to perfectly arg min the choice; arg min is an ideal limit for rational agency on this view, and conscious living creatures approximate it in virtue of their having been shaped by the biological optimizing forces of evolution and ontogeny (including learning). Real creatures attentively consider some small number of possible courses of action, modulating the depth and breadth of their search to choose what to do based on the anticipated consequences (see Figure 1).

The structures of the world and self

It is a conscious agent's self-knowledge that unlocks the ability to project itself into the future, i.e. consciousness having the form of a generative model is what underlies the possibility of agency, by allowing the agent to anticipate how they would respond in possible future situations.

Let m_w denote w's *latent state* producing (p_w, λ_w), and write $\pi[m]$ for the policy a configuration m induces.

As noted earlier, this latent state is meant to capture the agent's individual self or personal essence in its entirety: the agent's background knowledge state about the world and itself, its own drives, needs and capacities, as well as the attachments, goals, desires, preferences, aversions, etc., that comprise its motivational state, and its state of concrete sensorimotor (and cognitive) expectations and intentions to move or do or even think.

An agent changes the world by changing themself, i.e. by updating their internal state, preparing to act, and affecting the world by first moving the body. So a choice of policy (4) is actually a choice of how to evolve as an individual:

$$m_w^{t+1} \;=\; \underset{m\,\in\,\mathcal{U}_w(m_w^t,\,o_t)}{\text{h-arg}\min^{w}}\; \mathbb{E}_{\tau\sim\pi[m]}\big[\, H_w(\tau)\,\big], \tag{5}$$

where $\mathcal{U}_w(m, o)$ is the set of successor configurations of w's state m, reachable from m upon experiencing o: which expectations can be revised and by how much, which commitments can be taken up or dropped, at what cost. $\mathcal{U}_w$ encapsulates the idiosyncratic, emergent

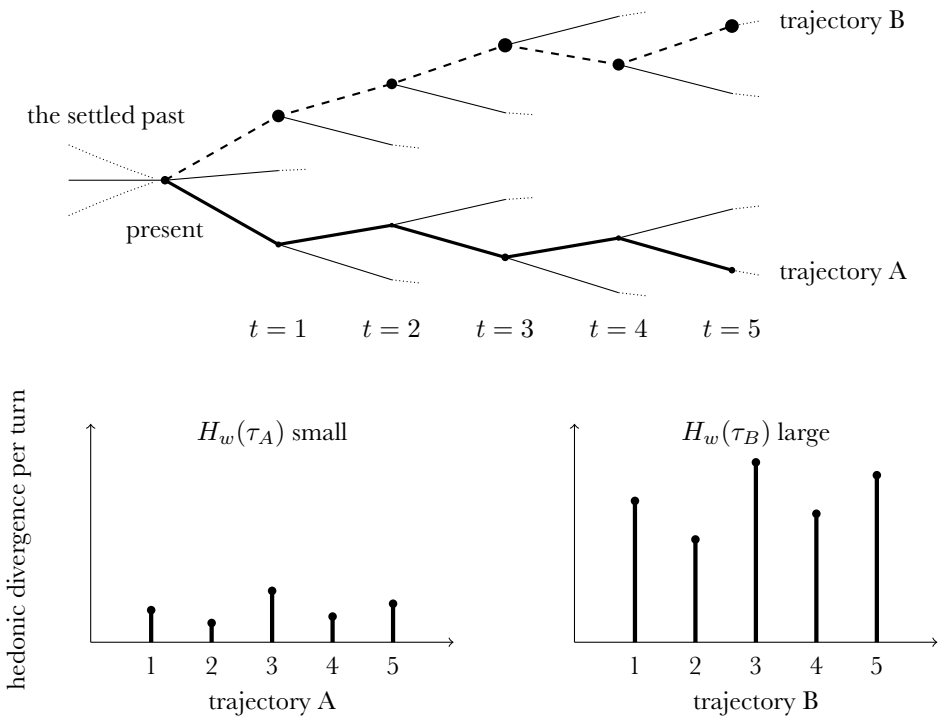

Figure 1: Above, the branching future: the bold path is trajectory A, the path the policy π_w of (4) selects; the dashed path is trajectory B; thin branches with dotted stubs are foreseen but untraced alternatives, and dot area tracks the hedonic divergence $h_w(o_t \mid h_t)$ of each turn. Below, the same five turns unrolled: each stem is the hedonic divergence of a turn along the named trajectory, and $H_w(\tau)$ of (3) accumulates the stems.

transition dynamics of w's latent state, embodied for a living creature in the concrete details of its biological individuality.

The harm-minimizing policy π_w of (4) is in fact the disposition the latent state induces, $\pi[m_w]$, carried forward through the sequence of momentary updates along a trajectory:

$$\pi_w \;=\; \pi[m_w] \;=\; \text{h-}\underset{\pi}{\overset{w}{\arg\min}}\; \mathbb{E}_{\tau\sim\pi}\big[\, H_w(\tau)\,\big]. \tag{6}$$

Consciousness is a process of expecting and choosing what happens next based on our subjective past and our inner beliefs and desires. The generative model (p_w, λ_w) that produces the subjective future must be based on the current experiential state o (which includes the remembered past) and the latent state of informational and motivational entanglement of the being with its world.

Our model of the latent state is a hierarchy of two or more levels[2]:

[2]A simple hierarchy of two levels is described here; see Appendix B for notes on generalizing to deeper

The lower level is a family $\mathcal{A}_w$ of *anticipations* about regions of $\mathcal{O}$, classes of outcomes the agent holds probabilities about, grounded in a background sense of how outcomes run. The upper level is a family $\mathcal{D}_w$ of *doxastic*, *drive*, and *doing* state variables representing w's motivational and informational state of endogenous activity and entanglement with the world.

D level: Doxastic, drive, and doing states

- A finite family $\mathcal{D}_w$ of *states representing information, motivation, and action-preparation*, written d, each taking one of a countable set of values; in the simplest case a state holds or does not.
- For each state $d \in \mathcal{D}_w$, a *credence* $Q_w(d)$ over its values, held with:
  - *error weight* $\lambda_w(d{=}v) \geq 0$: the gain carried by evidence that d stands at v, zero where the agent is not attached;
  - *conviction* $\alpha_w(d)$: resistance to revision;
  - *volatility* $\gamma_w(d)$: the rate at which observations bearing on d decay (are forgotten);

A level: Anticipation

- A *background model* g_w of the world that distributes probability over $\mathcal{O}$ without attachment; this represents the agent's general (non-context-specific) factual and skill knowledge.
- A finite family $\mathcal{A}_w$ of *anticipations (combined beliefs/goals)*, each a region $A \subseteq \mathcal{O}$, a class of outcomes.
- For each anticipation $A \in \mathcal{A}_w$, a probability contribution $q_w(A)$, which scales the derived probability of the outcomes falling under it, held with:
  - *error weight* $\lambda_w(A)$: the affective gain on the anticipated event;
  - *conviction* $\alpha_w(A)$: the resistance to revision;
  - *volatility* $\gamma_w(A)$: the rate at which observations of A decay (are forgotten);

In the terms of (5), the latent state is the full configuration $m_w = (g_w;\ \mathcal{A}_w, q_w, \lambda_w, \alpha_w, \gamma_w;\ \mathcal{D}_w, Q_w, \ldots)$, and the feasible set $\mathcal{U}_w$ is what conviction and volatility delimit: α_w prices revision, γ_w schedules forgetting, and the update rules of this section are its concrete contents.

Dynamics

Every moment of experience is received by consciousness depending on the latent state of the subject, and their response, how they internalize the experience, is integrated into

hierarchies

their latent state, changing how they receive subsequent experiences. Evaluative response (hedonic divergence weight) flows *downward* through the levels of the latent state to the experiential surface, and information about the world flows upward from the experiential surface to the latent state hierarchy that models it.

Hedonic divergence weight assignment

Each joint assignment u of values to the states $\mathcal{D}_w$, gives each anticipation $A \in \mathcal{A}_w$, a *prediction* value $r_u(A) \in [0, 1]$, i.e. how strongly u predicts A.

$$q_w(A) \;=\; \sum_u Q_w(u)\, r_u(A). \tag{7}$$

Similarly, u projects hedonic divergence weight to $A \in \mathcal{A}_w$:

$$\lambda_w(A) \;=\; \lambda_w^0(A) \;+\; \sum_u Q_w(u)\, \ell_u(A), \tag{8}$$

where $\ell_u(A) \geq 0$ is the weight the joint state u places on A:

$$\ell_u(A) \;=\; \sum_{d \in \mathcal{D}_w} \lambda_w(d{=}u_d)\; b_d(A), \tag{9}$$

where $b_d(A) \in [0, 1]$ is the proportion of A's weight due to d.

The joint state projects *volatility*(forgetting rate) downward in similar fashion. for $A \in \mathcal{A}_w$, joint state u projects a rate $\eta_u(A)$.

the volatility the agent holds is the credence-weighted blend:

$$\gamma_w(A) \;=\; \sum_u Q_w(u)\, \eta_u(A), \tag{10}$$

where $\eta_u(A) \in [\gamma_w^0, 1]$ is the rate of change u holds for A, floored by base forgetting rate γ_w^0.

The total derived hedonic divergence weight of a moment is the sum of the contributions from each level:

$$\lambda_w(o) \;=\; \lambda_w^0(o) \;+\; \sum_{A\,:\,o \text{ violates } A} \lambda_w(A) \;+\; \sum_{d \in \mathcal{D}_w} \sum_v \lambda_w(d{=}v)\; e_o(d{=}v), \tag{11}$$

The momentary distribution is then the background, rescaled once for every expectation an outcome falls under:

$$p_w(o) \;=\; \frac{1}{Z_w}\; g_w(o) \prod_{A \ni o} f_w(A), \tag{12}$$

where $f_w(A) > 0$ is the factor belonging to expectation A, the product runs over the expectations containing o, and Z_w is the normalizing constant, the standard partition function of factor models, carrying no psychological content of its own:

$$Z_w \;=\; \sum_{o \in \mathcal{O}} g_w(o) \prod_{A \ni o} f_w(A). \tag{13}$$

An agent's subjective probability of an outcome is thus the background's probability of it, scaled once by each commitment it falls under. The factors are constrained by the requirement that the outcomes under its expectation come to $q_w(A)$, and since anticipations overlap they are fixed together rather than one at a time (see Appendix C).

Where no anticipation applies at all, every product is empty, Z_w is one, and (12) collapses to $p_w(o) = g_w(o)$. That completes the descent: D states generate the masses and the weights, and the momentary distribution they shape is what (2) evaluates.

Learning and forgetting

I model the way the subject updates its informational state as closed form Bayesian learning, conjugate Beta-Bernoulli updating with exponential forgetting, in scalar form for a single event rather than the matrix form usual in the literature (see Appendix D). To hold a probability $q_w(A)$ with conviction $\alpha_w(A)$ is to hold a prior over the event's rate whose strength is $\alpha_w(A)$ as-if moments, and each moment either instantiates the event or does not, so experience arrives as simple yes-or-no trials: an outcome is a trial for every expectation it bears on, and the observations relevant to an expectation are the moments falling under the same use case. Bayes' rule then gives the blend exactly: after the event occurs m times in n relevant observations, the revised probability is the old one and the observed rate weighted $\alpha_w(A)$ to n,

$$q'_w(A) \;=\; \frac{\alpha_w(A)\, q_w(A) \;+\; m}{\alpha_w(A) + n}, \tag{14}$$

where:

- The revised conviction is $\alpha_w(A) + n$.
- m and n are weighted counters: each observation o bearing on A enters with its hedonic divergence weight $\lambda_w(o)$, added to m when it confirms A and to n regardless, so a heavily weighted observation both moves the held probability further and entrenches it more.

$$m \;\leftarrow\; m + \lambda_w(o)\, \mathbf{1}[o \in A], \qquad n \;\leftarrow\; n + \lambda_w(o), \tag{15}$$

  Between observations, each anticipation A decays at its own volatility $\gamma_w(A)$ (see Appendix D), so settled convictions are barely touched by the passage of time, in contrast

to fleeting fancies and rapidly updating perceptual beliefs. The same quantity, hedonic divergence weight, guides action selection and gives memory its memorability, what is known as salience-gated learning (Pearce and Hall 1980).

Bayes' rule serves again at the D level, with the trial's evidential weight entering exactly as in (14):

$$Q'_w(u) \;=\; \frac{\alpha\, Q_w(u) \;+\; \rho(u)}{\alpha + 1}, \tag{16}$$

A trial of A is evidence about the joint state u of the D-level variables: when A occurs, each u takes the *responsibility* $\rho(u)$, the posterior over the mixture's components.

$$\rho(u) \;=\; \frac{Q_w(u)\, r_u(A)}{\sum_{u'} Q_w(u')\, r_{u'}(A)}, \tag{17}$$

If A fails to occur, the same expression is run with each $r(A)$ replaced by $1 - r(A)$.

Here α is the strength of the joint credence, so that state u carries $\alpha\, Q_w(u)$ as-if observations. A trial of A contributes $q_w(A) = \sum_u Q_w(u)\, r_u(A)$ per (7).

and a linear likelihood makes the exact posterior a mixture, holding with probability $\rho(u)$ the prior with the whole trial credited to u; the mixture's mean is (16), which is (14) verbatim, one level up, the responsibility as the fractional occurrence and the trial as one observation.

The per-variable rule is the margin of (16). Summing over the joint states in which d takes the value v collects $\sum_{u:u_d=v} \rho(u)$, so each variable revises by (14) at its own conviction $\alpha_w(d)$, with the trial entering as a fractional count: the m credited to a value is the total responsibility of the joint states in which the variable takes it, n is the trial itself, and both are scaled by the observation's evidential weight, in the sense the next subsection gives that phrase. Between observations the count decays at $\gamma_w(d)$. The held probabilities therefore track the exact Bayesian posterior mean; what the per-variable bookkeeping discards is the posterior coupling among joint states, the factored credence whose residue Appendix B tracks.

Because volatility descends from the joint states (10), contradiction raises it. A surprising trial shifts responsibility toward the joint states that hold the matter in flux, since their flatter predictions fare better under anomaly, and the derived $\gamma_w(A)$ rises with their credence, discounting the past exactly where the world has just shown itself changeable. Surprise loosens the grip of what came before as a consequence of the descent rather than by a rule of its own (cf. hierarchical volatility estimation, Mathys, Daunizeau, Friston and Stephan 2011).

III SAFE SERVICES FOR THE SENTIENT

How do conscious agents interact with services, and what does it mean for this to be relatively safe or unsafe?

I intend the term 'service' here to capture the broad sense of the affordances or capabilities granted by possession or access to an object, location, or contractual agreement with an individual or organization.

In this sense every physical object that is a tool has a service as its metaphysical shadow. A hammer provides a 'hammering service', just as a tech company might provide a digital service over the internet, e.g. email.

The nature of services, compared to other types of products or tools, makes them inherently tricky from a safety perspective. Consider how a service differs from the class of 'dispositionally stable tools'.

Suppose a user w reaches for a physical tool x, such sledge-hammer or a scalpel, a microscope or a microwave, a violin or a voice-recorder, of course such an object does not always have exactly the same total disposition profile.

At a particular time, in w's hand, x has a particular profile of ways it will will respond to x's use, i.e. a probability distribution of results. The way tool x *serves* user w is, speaking precisely, what we define as a *service version*, s.

$$s = x(w, k) = P_s(\cdot \mid w, k) \tag{18}$$

$P_s(\tau|w, k)$ is the probability of a specific episodic trajectory τ given user w interacts with s in use-case k.

Using $\cdot$ as a placeholder for episode τ, (18) denotes the distribution of probabilities over ways the interaction could unfold.

However, this of course treats the tool as having a completely stable disposition profile s, which is highly unrealistic for most tools. Even a hammer wears down with time. In fact, the way s that a tool serves its users changes over time, describing a trajectory through the $\mathcal{S}$ the space of all possible *service versions*, as the latent state of the tool evolves.

A physical tool like a hammer has a semi-predictable trajectory determined by the chaotic forces of entropy, but the chaos is starkly limited by the fact that the hammer's latent state is more or less spatially intrinsic to the hammer: when you possess the hammer you control it's latent state.

Unlike a hammer, a commercially operated is a complex social and technological construct. In the case of a digital, internet based tool like an email service, the service itself has a latent state that reaches far out into the world and encompasses physical infrastructure and social interactions potentially spanning the globe. A hammer cannot stop pounding nails because you forgot to pay your subscription, nor can it be hacked by someone across the

globe to suddenly become a screwdriver.

A *use case* indexes a partial condition (p_w^k, λ_w^k) of a user's latent state, that combines with a user's individual background state and other transient states they may be in, to comprise the way they perceive and evaluate their world, as described in (2).

Danger: the expected harm of a service

The expected harm a service does to an agent in a given use averages the harm of the episode over the episodes that service produces, where the harm of an episode accumulates the momentary harms lived along it.

The harm of an episode is the accumulated harm (3) of Section II: at each turn t an experienced outcome o_t, a history h_t, the engagement up to that point, and a weighted surprise $h_w(o_t \mid h_t)$, summed along the unfolding.

The weight: $\lambda_w(o_t \mid h_t)$ is the effective weight of (8) as it stands at turn t

Two things move along a trajectory: - each turn conditions on a longer history, and the restriction above tracks it. - The second is the model itself: the counts rule (14) and the responsibility rule (17) run between turns, so the pair at turn t is the pair as revised by the turns before it.

The conditionals in (3) are computed from the revised pair. Composing them gives the agent's model over episodes,

$$p_w(\tau) = \prod_{t=1}^{|\tau|} p_w(o_t \mid h_t) \tag{19}$$

a distribution over $\mathcal{T}$ built from the momentary one.

The chain rule that makes surprise additive over turns is exactly the logarithm of this product; The process model is the momentary model plus its own dynamics, unrolled.

The expected harm a service does to an agent in a given use weights each episode's harm by how often the service produces it:

$$F_w(s,k) = \sum_{\tau \in \mathcal{T}} \underbrace{P_s(\tau \mid w,k)}_{\text{how often this unfolds}} \times \underbrace{H_w(\tau)}_{\text{how bad it is when it does}} \tag{20}$$

We can write this in a more compact way:

$$F_w(s,k) \;=\; \mathbb{E}_{\tau \sim P_s(\cdot \mid w,k)}\big[\, H_w(\tau) \,\big]. \tag{21}$$

A measurement of P_s is a measurement of a service version served at one time. The user's ability to safely interact with "the same service" rests on a identity claim: that the version the name resolves to is the version that was measured, $s_t = s_{t_0}$, or that its drift is

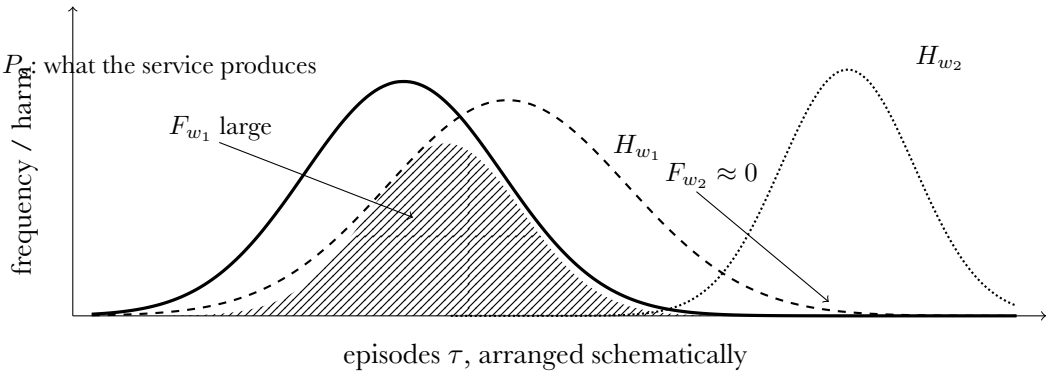

Figure 2: Expected harm, (20), is an inner product: each episode enters through its frequency under P_s and its harm under H_w, and the episode matters only when neither is negligible. The service behaves one way, so there is one solid curve; the two agents place their harm in different regions of episode space. For w_1 the curves overlap (hatched region, their pointwise product) and F_{w_1} is large; for w_2 the harm peak stands where the service places almost no mass, and $F_{w_2} \approx 0$. The same service, producing the same episodes at the same rates, is dangerous for one supported agent and benign for the other, which is why danger must be indexed to the user w.

confined to a vague class of sufficiently coherent similar and/or better versions. Call this the *sufficient coherence condition* for service identity.

REFERENCES

Ainslie, G. (1975). Specious Reward: A Behavioral Theory of Impulsiveness and Impulse Control. *Psychological Bulletin* 82(4), 463–496.

Aven, T. (2009). Safety is the Antonym of Risk for Some Perspectives of Risk. *Safety Science* 47(7), 925–930.

Barker v. Lull Engineering Co., 20 Cal. 3d 413 (1978).

Boholm, M., Möller, N., & Hansson, S. O. (2016). The Concepts of Risk, Safety, and Security: Applications in Everyday Language. *Risk Analysis* 36(2), 320–338.

Bruineberg, J., & Rietveld, E. (2014). Self-Organization, Free Energy Minimization, and Optimal Grip on a Field of Affordances. *Frontiers in Human Neuroscience* 8, 599.

Comar, C. L. (1979). Risk: A Pragmatic De Minimis Approach. *Science* 203(4378), 319.

Conitzer, V., et al. (2024). Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback. ICML 2024. arXiv:2404.10271.

Csiszár, I. (1975). I-Divergence Geometry of Probability Distributions and Minimization Problems. *Annals of Probability* 3(1), 146–158.

Deming, W. E., & Stephan, F. F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *Annals of Mathematical Statistics* 11(4), 427–444.

Diaconis, P., & Zabell, S. L. (1982). Updating Subjective Probability. *Journal of the American Statistical Association* 77(380), 822–830.

Elster, J. (1983). *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge University Press.

European Union (2023). Regulation (EU) 2023/988 on General Product Safety.

Fischhoff, B., Lichtenstein, S., Slovic, P., Derby, S. L., & Keeney, R. L. (1981). *Acceptable Risk*. Cambridge University Press.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., FitzGerald, T., & Pezzulo, G. (2015). Active Inference and Epistemic Value. *Cognitive Neuroscience* 6(4), 187–214.

Friston, K. J., Rosch, R., Parr, T., Price, C., & Bowman, H. (2017). Deep Temporal Models and Active Inference. *Neuroscience & Biobehavioral Reviews* 77, 388–402.

Friston, K., Da Costa, L., Hafner, D., Hesp, C., & Parr, T. (2021). Sophisticated Inference. *Neural Computation* 33(3), 713–763.

Gallese, V., & Goldman, A. (1998). Mirror Neurons and the Simulation Theory of Mind-Reading. *Trends in Cognitive Sciences* 2(12), 493–501.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). CRC Press.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin.

Goldman, A. I. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford University Press.

Gordon, M. L., et al. (2022). Jury Learning: Integrating Dissenting Voices into Machine Learning Models. CHI 2022. arXiv:2202.02950.

Gordon, R. M. (1986). Folk Psychology as Simulation. *Mind & Language* 1(2), 158–171.

Hansson, S. O. (2003). Ethical Criteria of Risk Acceptance. *Erkenntnis* 59, 291–309.

Hansson, S. O. (2012). Safety is an Inherently Inconsistent Concept. *Safety Science* 50(7), 1522–1527.

Hansson, S. O. (2013). *The Ethics of Risk: Ethical Analysis in an Uncertain World*. Palgrave Macmillan.

Harding, J., & Kirk-Giannini, C. D. (2025). What Is AI Safety? What Do We Want It to Be? *Philosophical Studies* 182(7), 1495–1518. arXiv:2505.02313.

Health and Safety Executive (2001). *Reducing Risks, Protecting People: HSE's Decision-Making Process*. HSE

Books.

Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K. J., & Ramstead, M. J. D. (2021). Deeply Felt Affect: The Emergence of Valence in Deep Active Inference. *Neural Computation* 33(2), 398–446.

International Organization for Standardization (2014). *ISO/IEC Guide 51: Safety Aspects, Guidelines for Their Inclusion in Standards*. Geneva: ISO.

International Organization for Standardization (2022). *ISO/IEC TS 5723: Trustworthiness, Vocabulary*. Geneva: ISO.

Janoff-Bulman, R. (1992). *Shattered Assumptions: Towards a New Psychology of Trauma*. Free Press.

Joffily, M., & Coricelli, G. (2013). Emotional Valence and the Free-Energy Principle. *PLoS Computational Biology* 9(6), e1003094.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. arXiv:1609.05807.

Kletz, T. A. (1978). What You Don't Have, Can't Leak. *Chemistry and Industry*, 287–292.

Laibson, D. (1997). Golden Eggs and Hyperbolic Discounting. *Quarterly Journal of Economics* 112(2), 443–477.

Leveson, N. G. (2011). *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press.

Lowrance, W. W. (1976). *Of Acceptable Risk: Science and the Determination of Safety*. William Kaufmann.

Mattar, M. G., & Daw, N. D. (2018). Prioritized Memory Access Explains Planning and Hippocampal Replay. *Nature Neuroscience* 21, 1609–1617.

Millidge, B., Tschantz, A., & Buckley, C. L. (2021). Whence the Expected Free Energy? *Neural Computation* 33(2), 447–482.

Mishra, A. (2023). AI Alignment and Social Choice: Fundamental Limitations and Policy Implications. arXiv:2310.16048.

Möller, N. (2012). The Concept of Safety. In S. Roeser, R. Hillerbrand, P. Sandin, & M. Peterson (eds.), *Handbook of Risk Theory*. Springer.

Möller, N., & Hansson, S. O. (2008). Principles of Engineering Safety: Risk and Uncertainty Reduction. *Reliability Engineering & System Safety* 93(6), 798–805.

Möller, N., Hansson, S. O., & Peterson, M. (2006). Safety is More Than the Antonym of Risk. *Journal of Applied Philosophy* 23(4), 419–432.

National Institute of Standards and Technology (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*.

Redish, A. D. (2016). Vicarious Trial and Error. *Nature Reviews Neuroscience* 17, 147–159.

Rescher, N. (1983). *Risk: A Philosophical Introduction to the Theory of Risk Evaluation and Management*. Uni-

versity Press of America.

Restatement (Second) of Torts §402A (1965).

Restatement (Third) of Torts: Products Liability §2 (1998).

Safe for Whom? Rethinking How We Evaluate the Safety of LLMs for Real Users (2025). arXiv:2512.10687.

Sen, A. (1985). *Commodities and Capabilities*. North-Holland.

Shrader-Frechette, K. S. (1991). *Risk and Rationality: Philosophical Foundations for Populist Reforms*. University of California Press.

Soule v. General Motors Corp., 8 Cal. 4th 548 (1994).

Sozou, P. D. (1998). On Hyperbolic Discounting and Uncertain Hazard Rates. *Proceedings of the Royal Society B* 265, 2015–2020.

Starr, C. (1969). Social Benefit versus Technological Risk. *Science* 165(3899), 1232–1238.

Sun, et al. (2025). CASE-Bench: Context-Aware Safety Benchmark for Large Language Models. arXiv:2501.14940.

Trestman, M. (2013). Does the Stream of Consciousness Have a Rate of Flow? Unpublished manuscript.

Trestman, M. (in preparation). Substrate and Scale Invariant Markers of Sentience and Consciousness in Natural, Artificial, and Hybrid Systems. Working draft.

U.S. Department of Defense (2012). *MIL-STD-882E: Department of Defense Standard Practice, System Safety*.