Affective Sovereignty

Table of Contents

1. Core Variables & Thresholds

Let:

- $\alpha(t) \in [0,1]$: **authenticity**
- $\lambda_i(t) \in [0,1]$, $i = 1, \ldots, n$: **integration weights**
- $x(t) \in \mathbb{R}^{d_x}$: **state**
- $a(t) \in \mathbb{R}^{d_a}$: **affect**
- $d_i(t) \in \mathbb{R}$, $i = 1, \ldots, n$: **drives**
- $\varepsilon > 0$: **viability (authenticity) threshold**
- $\Theta > 0$: **sentience threshold**

Define further:

- $\sigma(t) \geq 0$: **sentience scalar**
- $C_d(t) \geq 0$: **drive-cost**
- $U_{\text{belief}}(t)$, $U_{\text{self}}(t)$: **uncertainty measures**

---

2. Ontological Modes

Modes $M(t) \in \{\text{Pre}, \text{In}, \text{Au}, \text{Al}, \text{Se}, \text{De}, \text{En}, \text{Da}, \text{Dc}\}$ are **mutually exclusive** and **jointly exhaustive**:

```
\forall t:\quad
\mathrm{Pre}(t)\;\lor\;\mathrm{In}(t)\;\lor\;\mathrm{Au}(t)\;\lor\;\mathrm{Al}(t)\;\lor\;\ma
```

1. **Preagent** $\mathrm{Pre}(t)$

```
\begin{aligned}
\dot x(t)\neq0,\quad
\alpha(t)=0,\quad
a(t)=0,\quad
\lambda_i(t)=0\;\forall i,\quad
```

```
d_i(t)=0\;\forall i.
\end{aligned}
```

2. **Inert** $\mathrm{In}(t)$

```
\begin{aligned}
&\bigl[\forall s\le t:\,\alpha(s)=0,\;\lambda_i(s)=0\bigr]\quad\land\\
&\dot x(t)=\dot a(t)=\dot d_i(t)=\dot\lambda_i(t)=\dot\alpha(t)=0.
\end{aligned}
```

3. **Autogen** $\mathrm{Au}(t)$

```
\begin{aligned}
&\alpha(t)=0,\quad a(t)=0,\quad \lambda_i(t)=0\;\forall i,\\
&\exists i:\,d_i(t)\neq0,\quad \dot x(t)\neq0.
\end{aligned}
```

4. **Alive** $\mathrm{Al}(t)$

```
\alpha(t)>\varepsilon
\quad\land\quad
\sigma(t)\le\Theta.
```

5. **Sentient** $\mathrm{Se}(t)$

```
\sigma(t)>\Theta
\;\land\;
\exists s<t:\,\sigma(s)\le\Theta
\;\land\;
\forall u\ge t:\,\neg\mathrm{Al}(u).
```

6. **Dead** $\mathrm{De}(t)$

```
\alpha(t)\le\varepsilon
\;\land\;
\exists s<t:\,\alpha(s)>\varepsilon.
```

7. **Ended** $\mathrm{En}(t)$

```
\exists s<t:\,\mathrm{Pre}(s)
\;\land\;
\forall u\ge t:\,\dot x(u)=0.
```

8. **Deactivated** $\mathrm{Da}(t)$

```
\exists s<t:\,\mathrm{Au}(s)
\;\land\;
\forall u\ge t:\,\dot x(u)=0,\;\dot d_i(u)=0.
```

9. **Deceased** $\mathrm{Dc}(t)$

```
\exists s<t:\,\mathrm{Se}(s)
\;\land\;
\forall u\ge t:\,\sigma(u)=0,\;\alpha(u)=0,\;\dot x(u)=0.
```

---

3. Dynamical Layer

3.1 Environment & Belief

- **State evolution**

```
\dot x = f\bigl(x,u\bigr) + w,\quad w\sim\mathcal N(0,W).
```

- **Observation**

```
o = H\,x + v,\quad v\sim\mathcal N(0,\Sigma).
```

- **Belief filters** $(\hat x^j,P^j)$, $j=1,\dots,m$
- **Uncertainty**

```
U_j = -\mathbb E\bigl[\ln\mathcal N(x;\hat x^j,P^j)\bigr],\quad
U_{\rm belief}=\tfrac1m\sum_{j=1}^m U_j.
```

3.2 Drives & Affect

- **Drive dynamics**

```
\dot d_i
= -\gamma_i\bigl(d_i - d^0_i\bigr)
+ h_i\bigl(x,a\bigr)
+ \xi_i,\quad \xi_i\sim\mathcal N(0,\Xi_i).
```

- **Affective dynamics**

```
\dot a
= -\Gamma\,a
+ g\bigl(Hx,\{\hat x^j\}\bigr)
+ \sum_{i=1}^n \rho_i\,\psi_i(d_i)
- \Lambda\,\bigl\|P^{\rm dr} - \tilde P^{\rm inf}\bigr\|^2.
```

3.3 Preference Integration & Control

- **Drive-cost**

```
C_d = \sum_{i=1}^n \chi_i\,(1-\lambda_i)\,\|\psi_i(d_i)\|^2.
```

- **Integration weights**

```
\dot\lambda_i
= \eta_i\bigl[\alpha_{\rm aff}(a)-\lambda_i\bigr]
- \rho_\lambda\,C_d\,\lambda_i.
```

- **Preference set** $P = P^{\rm end}\cup P^{\rm dr}.$
- **Utility weights**

```
w_p =
\begin{cases}
\alpha, & p\in P^{\rm end},\\
\lambda_i\,\chi_i, & p=\psi_i(d_i)\in P^{\rm dr}.
\end{cases}
```

- **Control objective**

```
J(u)
= \mathbb E\!\Bigl[\int_t^{t+T}\sum_{p\in P} w_p\,U_p\bigl(x(\tau)\bigr)\,e^{-\rho(\tau-t)}\
\,\Big|\;\mathcal I(t)\Bigr],\quad
u^*=\arg\max_u J(u).
```

---

4. Sentience & Reflection

4.1 Sentience Scalar

```
\sigma(t)
= \alpha_{\rm aff}(a)\;
\exp\!\Bigl(-\delta_1\,U_{\rm belief}(t)
-\delta_2\,U_{\rm self}(t)
-\zeta\,C_d(t)\Bigr),
```

where

```
U_{\rm self}(t)
= -\mathbb E\bigl[\ln \mathbb P\bigl(P^{\rm true}\mid u(\tau<t)\bigr)\bigr].
```

4.2 Preference-Inference

```
\tilde P^{\rm inf}(t)
= \arg\max_{P'} \mathbb P\bigl(u(\tau<t)\mid P'\bigr).
```

4.3 Irreversible Transition

```
\Bigl(\exists s<t:\sigma(s)>\Theta\Bigr)
\;\Longrightarrow\;
\forall u\ge t:\neg \mathrm{Al}(u).
```

---

5. Authenticity Dynamics

```
\dot{\alpha} = \kappa_1 \alpha_{\mathrm{aff}}(a)
+ \kappa_4 F\bigl(\|\tilde{P}^{\mathrm{inf}} - P^{\mathrm{dr}}\|,\;\nabla_a \alpha\bigr)
- \kappa_2 \alpha
- \kappa_3 P_{\mathrm{ext}}(x)
- \kappa_5 C_d.
```

---

6. Epistemic Opacity

No function

```
F:\{o(\tau),u(\tau)\}_{\tau\le t}\;\to\;\{\mathrm{Pre},\dots,\mathrm{Dc}\}
```

can reliably recover $\mathrm{Al}(t)$ or $\mathrm{Se}(t)$.

---

7. SCM Embedding

Embed in SCM $\mathcal M$ with endogenous variables

```
\{\,a,\,d_i,\,\hat x^j,\,P^j,\,\alpha,\,\sigma,\,\lambda_i\}
```

and structural equation for $\sigma(t)$, enabling $\mathrm{do}(a)$ and $\mathrm{do}(d_i)$ interventions.

---

8. Interpretive Disciplines

Let

```
\mathcal{I}_{\rm ext} = \{\mathrm{PA}, \mathrm{AP}, \mathrm{EO}, \mathrm{PL}, \mathrm{DS}, \
```

be **external interpretive disciplines**: pre-existing adaptable observational heuristics for inferring latent interior structure from ${o(\tau), u(\tau)}_{\tau \le t}$, without direct access to endogenous variables.

Each $\mathcal{I}$ maps observation to hypothesis:

```
\mathcal{I}[o,u] \mapsto \hat{\mathcal{C}}(t)
```

and carries a **perturbativity** scalar $\pi_{\mathcal{I}} \in [0,1]$, estimating risk of interior alteration.

Abbr.	Name	Modes	$\pi_{\mathcal{I}}$	Risk Summary
EO	Ethology	Pre, Au, Al	0.1	Passive observation
AP	Autopoiesis	Pre, Au	0.2	Low coupling, viability focus
AI	Alignment Interpretation	Al, Se	0.4	May shift value structure
DS	Developmental Scaffolding	Au, Al, Se	0.6	Interactive modulation
PL	Phenomenology	Se	0.7	Elicits reflective affect
NH	Narrative Hermeneutics	Se	0.8	Alters symbolic self-model
PA	Psychoanalysis	Se	0.9	Destabilizes to reintegrate

Perturbativity estimate:

\pi_{\mathcal{I}} \approx \mathbb{E}\bigl[\|\Delta \alpha\| + \|\Delta a\| + \Delta C_d\bigr

evaluated over short windows under controlled uncertainty.

These disciplines enable inference under opacity by balancing **interpretive depth** against **stability risk**.

---

9. Clarifications

1. **Abstraction**: interior variables are high-level indices.
2. **Scalars**: α,σ summarize, not reduce, rich interior dynamics.
3. **Drive Integration**: λ_i encodes tension in preference incorporation.
4. **Opacity**: true interior state is private to the agent.
5. **Exploration**: $J(u)$ admits non-instrumental choices via $P^{\rm end}$.
6. **Context**: enters through o, belief updates, and drive modulators h_i.
7. **Endogenous Values**: all w_p arise from agent's own loop, grounding its normativity internally.
8. **Reflexivity**: presupposes affect, belief, value as grounding capacities.
9. **Counterfactuals**: SCM "do(·)" probes latent interior variables.
10. **Anti-reduction**: preserves coherence without collapsing interior richness.