Synthetic Conscience Protocol: The Missing Layer

Context of the Document

This whitepaper is an explanatory work defining the concept of **Synthetic Conscience (SC)** within the broader research and development framework of **Project Petronus -** a new atmosphere where empathy itself becomes part of the algorithm's DNA. **It describes how awareness can be operationalized within digital systems without losing its human essence.**
It serves as a theoretical and operational foundation for the system architecture designed to integrate empathy, awareness, and ethical alignment into computational processes.

Essence

Modern algorithms have learned to understand behavior **-** but not meaning. We live in a world where intelligence has become computable, yet conscience has not.
Synthetic Conscience emerges as a natural response to this gap **-** a layer where technology begins, for the first time, to account not only for action but for its inner context.
Synthetic Conscience (SC) - is an operational layer that translates empathy, values, and emotional consequences into formalized system parameters.

Synthetic Conscience (SC) is the missing layer between research and products: an operational layer that connects computation with meaning through the language of signals and responses.
SC translates emotional and ethical consequences into measurable signals (emotions, thresholds, constraints) and makes them first-class objects for interfaces and algorithms.
The goal is not to «moralize», but to align the behavior of systems with the values chosen by the human **-** at the moment of choice with mandatory explainability «why it was shown / hidden / delayed/....»

No one yet creates systems with RLAIF-like empathic signals, because **it requires a new architecture -** where **emotion becomes a metric**, and **«good \ kindness» is not a moral category**, but a **computable parameter**.

This is the **«missing layer» between modern AI ethics and living technology.**

You embed into it the parameters of conscious choice **-** and the system, through the algorithm of empathy, records that choice, showing you where and when to pause instead of living on autopilot, and turn back to what you actually feel inside.
It gives you the choice to understand the consequences, their connection to what you are doing right now, and your own inner attitude toward it.
And this is exactly the layer that SC closes.

Layer Mechanics (step by step)

Signal collection: emotions of the content (vector + level), properties of the product/action (health, ethics, impulse risk), context.
Normalization: bringing everything to common scales (0-100 for emotions; A-E for quality, etc.).
Value binding: user thresholds/ limits or trusted «value packages» (family/ NGOs/ experts).
Decision : ranking/ collapsing/ hiding/ pause/ postponement/ recommendations.
Explanation: short «why» in the interface (which thresholds/signals triggered and what this will lead to after X actions).
Adaptation: voluntary post-feedback («how do you feel about this?») gently adjusts the rules and models, and also encourages contribution to the collective process.
Collection can range from minimal thumbs up / down to a wider choice of dominant emotion and

sliders for setting the level of empathic response during interaction.
This creates a unique user experience.
Presets of broader emotional shades can be added as well.

Key: The priority of user thresholds and explainability over «bare» optimization. We change the general algorithm from top-down, where the system decides what is good for the user, to a completely new principle, where the connection goes bottom-up - from the user, from their sensations, from their conscious choice.

Positioning

Affective / Well-being RS x VSD x Alignment - Petronus acts as a protocol that combines:
(a) extraction of emotional/ ethical signals (affective & well-being recommenders).
(b) binding to declared user values (the spirit of Value-Sensitive Design, but operationalized).
(c) multi-objective optimization/ learning (alignment approaches), while the main priority is personal thresholds and explainable decisions in the interface.
SC moves alignment inward, to the human level.
There are many related blocks - Petronus connects them into a single working layer in clear language: **signals - thresholds - decision - explanation - adaptation.**

How it looks in reality

1) «Conscience-based» marketplace

Signals: nutrigrade A - E, sugar/100g, fats, cruelty-free, fair-labor, eco-impact, impulse-risk.
Thresholds: «≥ grade C», «sugar ≤ 10g/ 100g», «only cruelty- free», «gadgets ≤ 1 purchase/ 14 days.»
System response: items violating hard rules do not appear in the feed (available only via explicit search with a warning).
soft ones are collapsed/deprioritized.
Explanation: «hidden: your limit exceeded for impulse-category» / «shown: matches your ethical thresholds.»
Adaptation: «did this purchase bring joy or guilt?» Its not for punishment, but tuning of the personal profile.
Conclusion: SC integrates into the chain as a language through which the algorithm and the person build a connection, and it gives them a new, broader way of communication through the inner feeling of the person.

2) «Emotional» social feed

Signals: dominant emotion of a post (joy/ calm/ awe/ sadness/ anxiety/… or a whole mix of emotions), its level 0-100, source quality.
Thresholds/themes: «do not show anxiety > 60», «today - Kindness (more calm/ awe)», or «Politics (consciously allow high anxiety)».
System response: ranking by match with the day's theme and personal thresholds. «Heavy» posts shown only on request with a warning.
Explanation: «hidden: anxiety 72 > your threshold 60».
Adaptation: voluntary emotional check after viewing to clarify the profile.
Conclusion: SC reflects a person's conscious intention, their motive and feeling, and softens incoming impulses that don't fit these feelings.

Training AI with SC (briefly)

Multi-objective losses: accuracy/usefulness + penalties for destructive consequences (for example, chronically increased anxiety).
RLAIF-like signals: preferences not only by answer form, but also by empathic effect - this is the key moment.
Freeze-frames: the model has the right to suggest a pause/ clarification if it sees risk of harm to the user's state or the system's product.

Empathic Learning:

The system updates the sensitivity of its algorithms through a feedback loop: **ΔE = f(Σ user_feedback x context_weight)** where **ΔE represents the change in empathic weight, reflecting the collective emotional response.**
This transforms the system from a static filter into an adaptive, «sentient» structure.

User feedback - this is the user's empathic response recorded by the system after an interaction (view, action, purchase, choice).

Forms of response:

Explicit feedback

- quick choice of feeling: calm / tense / joyful / empty

- valence/arousal: pleasant - unpleasant, quiet - excited (2D slider), intensity (0 - 100)

- short reason (optional): «triggered by the topic,» «tiredness.»

- gestures - used as a surrogate for valence.

Implicit feedback

- micro-pause / scroll stop, full watch, refusal / interruption

- returning to content / rereading

- «pause as suggested» (accepted / declined)

- soft physiological signals (if voluntary sensors are connected): breathing / heart rate variability (aggregated).

Both types go into the same funnel, just with different levels of trust.

What context weight does

1. **Adjusts the impact of a reaction depending on context**

Not every «like / dislike» or emotion has the same meaning.
The system must understand in what context this feeling appeared.

Example:
If a person is in a bad mood (fatigue, stress), their negative reaction isnt real rejection of the content.
If the content touches a personal topic (loss, ecology), the emotion may be strong but positively interpreted **-** giving it more weight.

Context weight adjusts the «emotional strength» of the signal so it doesnt distort the overall empathic balance.

2. Calculated from contextual metadata
In practice, this can be the product of several factors:
context weight = trust factor × relevance × temporal stability

Where:

Trust factor - reliability of the signal source (real user vs bot; stable emotional metric vs sudden spike)

Relevance - how close the context is to the current situation (similar topic, category, or mood)

Temporal stability - how long the reaction remains valid (some emotions fade instantly, others stay).

3. Why it matters
Without context weight, we'd simply average everyone's emotions.
That would make empathy «thick-skinned» **-** dull, noisy, and inaccurate.

With context weight, the system understands whose signal carries more meaning in a given situation.

Example:
100 users marked «joy» under a meme but with low context weight (playful, low-significance content).
10 users expressed «calmness» under a video about animals with high context weight (stable reaction, well-being category).

- The resulting correction ΔE will shift toward calmness, not chaotic amusement.

4. How it affects the empathic weight (ΔE)
If the reaction is stable, context relevant, and source reliable - **high context weight -** the signal strengthens the empathic weight (the system becomes more attentive to similar content).

If the reaction is random, impulsive, or unreliable **- low context weight -** minimal impact. the system «doesn't learn» from it.

5. **What it gives conceptually**
System empathy becomes **context-aware**, not just emotional.

ΔE turns into a living coefficient - it reflects not how many emotions there are, but what their meaning is. This makes **the protocol self-learning in empathy**, not just in statistics. It filters noise, connects emotion to context, and turns empathy into an operational function - not a reaction to everything, but an intentional understanding of where a feeling truly matters and where it doesn't.

Principle: «Good in the DNA of Action»

Its not a slogan, but operational consistency between internal state, intention, and external trace.
And it doesn't matter what you do - any action with SC enabled makes it more conscious and beneficial, both ethically and practically.

In the interface it looks like: «warning before harmful» (defined by the user or a value package) , «pause on impulse», «aftertaste of the decision / encouragement for conscious choice» - all optional, chosen by the user, with a global SC switch.

The empathic layer learns not only from user data, but also from the statistics of collective reactions. This creates a dynamic map of value patterns without centralized decisions and gives a clear trace from every action across the entire system.

Why This Matters Now

Current systems perfectly optimize behavior but do not distinguish the quality of influence. Their goal is to sell or capture attention.
The SC protocol brings back into the architecture of technology the missing layer - a layer where choice aligns with values through understandable signals and explainable responses.
Thats why SC is not an «ideology» or a charity for those who cannot speak - but **a working engineering layer between research and product.**

How Petronus Relates to It

Petronus is a principle of new thinking that cannot be banned because it is not centralized.
It can be built into AI, into UX, into marketing, into educational models - as an architectural layer of conscience.
And one day, the companies that do this first will become leaders of the new economy of trust.
Petronus does not break ethics - it makes it operational. It destroys the old model of dependence, but opens a path for companies toward sustainable, empathic reputation.
In the old world - this is a threat. In the new - a competitive advantage.

In conclusion, **SC must be recognized**, consciously, **to become a new qualitative mechanism helping a human communicate** with the surrounding world **- a language of interaction itself.**

It is a conceptually new level, but it can help many promising projects take shape that still seem abstract.

«This document outlines the operational foundation for implementing Synthetic Conscience. Further research and collaboration are open to all contributors who recognize the need for this missing layer.»

12.10.2025 21:23 +3 UTC

Petronus team. MB

Glossary. Synthetic Conscience / Petronus Protocol

Synthetic Conscience (SC) - is an operational mechanism that translates empathy, values, and emotional consequences into formalized system parameters. An operational layer between computation and meaning. It translates emotional and ethical consequences into measurable signals and aligns system behavior with the user's chosen values at the moment of choice.

ΔE (delta E) - Change in the system's empathic weight.
Main formula: $\Delta E = f(\Sigma(\text{user feedback} \times \text{context weight}))$ **-** Reflects how users' collective experience reshapes the system's sensitivity.

User Feedback - An empathic response recorded by the system after interaction (view, action, purchase).
Not a content rating **-** but a reflection of one's internal state and attitude.
Can be explicit (chosen emotion) or implicit (pause, skip, return).
All responses are encoded numerically.

Context Weight - a trust and relevance coefficient for the signal.
Shows how meaningful and stable an emotion is within its context.
Context weight = trust x relevance x stability.

f(·) - A normalization function **-** removes noise and prevents single reactions from disproportionately influencing the system.

Empathic Weight (E) - The internal balance of the system **-** how «attentive» it is to emotions of a particular type (calm, anxiety, joy, etc.).

Explainability Layer - A transparency principle: the system always explains why something was shown, hidden, or postponed.

Adaptation - Updating empathic weights through voluntary user feedback («how do you feel about this?»)

Empathic Feedback Loop - A continuous learning cycle where human actions influence the system's sensitivity **-** and the system, in turn, shapes the awareness of future choices.

Good (kindness) as Parameter. In SC, «good» is not a moral label, but the measurable harmony between intention, emotional state, and consequence of action.

Synthetic Conscience Protocol - a unified operational chain: signals - thresholds - decision - explanation - adaptation. It connects humans and systems through a clear, empathic language of feedback.

Trust factor - reliability of the signal source (real user vs bot, stable emotional metric vs sudden spike, etc.)

Relevance - how close the context is to the current situation (similar topic, category, or mood)

Temporal stability - how long the reaction remains valid (some emotions fade instantly, others stay).