

AGOS — un OS agentico **in produzione.**

Un sistema che lavora da solo, e le scelte per potersene fidare.



DARIO MARTINELLI

E tu, come fai a saperlo?

Noi siamo vigilati Consob e Banca d'Italia

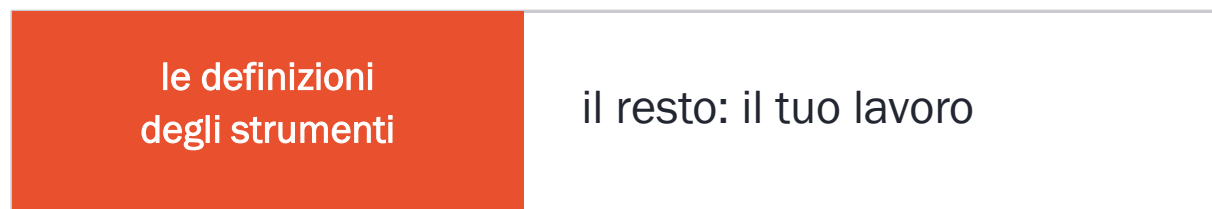
La piattaforma è costruita per **subire un audit da parte delle authority**.
Non esiste ancora una legge, ho usato il buon senso sapendo cosa controllano sulle persone.

Tracciatura di ogni evento e versioning di ogni modifica: chi ha fatto cosa e quando.

Se è un'azione schedulata: chi ha modificato l'automatismo e tracciatura di ogni esecuzione.

A maggio dicevo: il prompt "seziona" lo spazio semantico. Ecco il conto dei nostri tool.

UNA FINESTRA DI CONTESTO DA 32.000 TOKEN



42% della finestra

13.542 token di sole definizioni



e li paghi di nuovo a ogni singolo turno di conversazione

Le ancore sono i pesi fissi

Le definizioni degli strumenti, le istruzioni di sistema, i file già letti. Non li scrivi tu a ogni domanda: viaggiano con te, sempre.

Averle misurate cambia le scelte

Le definizioni degli strumenti occupano il 42% della finestra prima che l'utente abbia aperto bocca. Da qui in poi ogni decisione di questo sistema nasce da un numero, non da un'impressione.

Che cosa fa, quando qualcuno lo usa

```
chat · agentos.cuniberti.org

utente quanto valgono le case al metro quadro
in corso Francia, a Torino?

agos › cerca_zona("Torino, C.so Francia")
› prezzi_zona(id_zona)

2.288 €/m² media della zona (142 annunci)
```

le due righe rosse sono chiamate a codici python,
non testo generato da LLM

Chi lo usa

Undici persone, dal browser o da un bot Telegram.
Dall'altra parte non c'è un modello che ricorda: c'è
un sistema che va a cercare il dato e lo riporta con la
sua fonte.

E fa anche cose che nessuno gli chiede

Sorveglia sé stesso, ripara quello che sa riparare,
propone modifiche al proprio codice, e legge quello
che fanno gli altri.

14

settimane

786

commit

11

persone

81

strumenti

103.885

righe di test

Ma la cosa interessante è quella che nessuno gli chiede

IL BATTITO: SENZA CHE NESSUNO GLIELO CHIEDA

osserva	la coda di lavoro e gli errori
diagnostica	perché è successo
decide	quanto può rischiare in token
agisce	ripara, rigenera, propone
verifica	l'errore è davvero sparito?
registra	cosa ha imparato

se la verifica fallisce non chiude: riprova, o chiama un umano

Si sveglia su un evento

Quando arrivano richieste è il database a svegliarlo. Altrimenti a ogni giro segna l'ora in cui è passato: se in quell'ora smette di aggiornarsi, un guardiano orario esterno lo nota. (Hermes su Hetzner)

L'autonomia si guadagna

Rigenera gli strumenti rotti da solo. Tocca la produzione solo dopo una serie di successi verificati dai test.

Se un vincolo conta davvero, **non lo scrivi nel prompt**

Un'istruzione vale finché il modello se la ricorda. Un binario vale anche quando sbaglia.

Chi può leggere che cosa

sta nei permessi del database

non in una frase che dice all'agente di non guardare i dati degli altri

Quanto può spendere

sta nel codice con chiamate limitate

e se non riesce a sapere quanto ha già speso, non chiama

Dove può arrivare

nella macchina virtuale isolata dei test

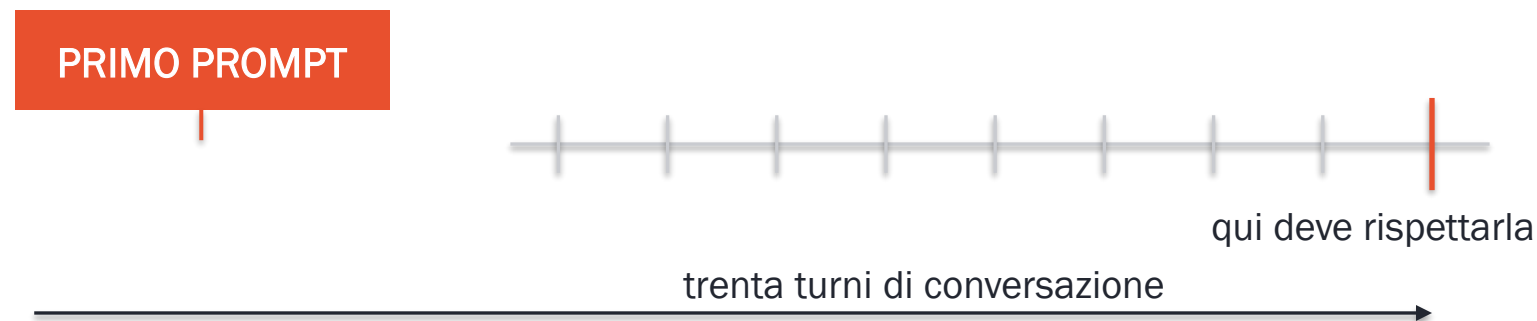
non nelle istruzioni che legge all'inizio

Il prompt resta la prima difesa. **Non è mai l'unica.**

La stessa regola, scritta in due posti diversi

«Non calcolare tu la volatilità: usa il valore che ti do io.»

SCRITTA NELLE ISTRUZIONI DI SISTEMA



27/30
violazioni

SCRITTA INSIEME AI DATI CHE RICEVE



0/30
violazioni

Anche il codice di questo sistema è scritto da agenti

Ma attraverso una macchina a stati, dove il passaggio di stato non lo decide l'agente che ha scritto.



```
review.json - il verdetto del revisore

"blocking": 2,
"findings": [ { "gravita": "major"
  "come_riprodurre": "..." } ]
```

Il verdetto è un file, non un'opinione

Il revisore non scrive «mi sembra a posto»: elenca i rilievi, con la riga e il modo di riprodurli. Finché i bloccanti non sono zero, lo stato non avanza.

Uno sprint vero: 19 iterazioni, 14 ore, 455.600 token, 3 review bocciate su 8 — **tutte fondate.**

« **Non ho guardato** »



Un controllo che non ha osservato non ha il diritto di emettere un verdetto.

Il controllo diceva: sono peggiorati i modelli.

Era falso.

Ogni domenica un controllo automatico misura se il sistema risponde bene con modelli diversi di Ollama Cloud. Un giorno ha bocciato tre modelli.

● ● ● sentinella · esito del giro settimanale		
gemma4	1/4	KO
glm-5.2	1/4	KO
deepseek	1/4	KO
tre modelli diversi, gli stessi tre casi		
strumenti inviati: 81 eseguiti: sì contatore: 0		

Tre modelli che sbagliano identici

non sono tre guasti diversi. E l'unico caso promosso era l'unico che non richiedeva alcuno degli 81 strumenti disponibili: il sospetto è partito da lì.

Gli strumenti giravano, era rotta la misurazione, non la cosa misurata.

Adesso il controllo sa dire: «non ho potuto misurare»

Due misure indipendenti della stessa cosa: probabilistico e deterministico. Quando non concordano, il verdetto non è «guasto».

le due misure concordano, e il caso passa

esito valido

le due misure concordano, e il caso fallisce

guasto vero

le due misure non concordano

non ho potuto misurare

Non è un test in più

È una spia deterministica messa nel punto obbligato di passaggio, che conta quello che passa davvero. Zero chiamate in più al modello.

E l'ho provato rompendolo

Contatore sabotato di proposito: «sei strumenti eseguiti, il contatore ne vede zero» → misura non attendibile.

Si ripara da solo. E si riscrive da solo.

Ma in quest'ordine: prima le rotaie, poi l'autonomia. Ecco com'è chiedere il permesso.

Telegram • AGOS → Dario

Ho preparato la modifica che avevi chiesto: suite verde, il revisore non ha rilievi.

Costo finora: 0,097 € su 2,00 €.

Approva

Rifiuta

run 3qjvwu6r — «Approva» premuto dal telefono

Si ripara

Una sessione verso un fornitore esterno moriva 26 volte in tre settimane, e ogni volta serviva un intervento umano. Ora lancia un segnale e si risveglia da sola: misurato in cinque secondi.

Si riscrive

Un agente scrive codice su questo stesso progetto: in una macchina virtuale isolata, con un revisore che cerca il difetto, un tetto di spesa che nel dubbio ferma tutto, e quel bottone lì a sinistra.

E legge quello che fanno gli altri, al posto mio

Tre dei cantieri feature di quest'anno sono nati dall'osservazione di repo furbi su GitHub.

01 · GUARDA

18 repository, ogni domenica

L'unità è la release, non il commit: sei su otto fanno più di cento commit al mese, e leggerli tutti è rumore.

02 · GIUDICA

Contro un profilo scritto

Non «il modello pensa che sia interessante», ma un confronto rispetto alla costituzione del mio progetto.

03 · PROPONE

Diventa una request

Non una notifica: entra nella stessa coda del lavoro umano.

Può anche dire «questa cosa non serve».

Non chiedere al sistema se funziona.

Chiedigli come fa a saperlo.

Un vincolo che conta è deterministico, non probabilistico.

L'ultima cosa che ricordi di controllare è lo strumento con cui controlli.

Splitvibe **Manager**

due persone, un conto corrente virtuale condiviso, otto anni di spese e nessuna voglia di classificarle.

Il primo settembre il fornitore ha chiuso la porta

OTTO ANNI DI STORICO DENTRO UN SERVIZIO CHE NON CONTROLLO

fino al 2026	Il mastro era Splitwise: comodo, gratis, e fuori dalle mie mani.
01/09/2026	Per usare l'API serve l'abbonamento: una decisione presa senza di me.
oggi	Una tabella mia: 1.818 spese, con gli identificativi originali.

La lezione, in una riga

La prima domanda su un fornitore non è quanto costa: è che cosa succede quando decide da solo di cambiare le policy?

Il criterio di riuscita non era «l'import non ha dato errori».

Era: il saldo fra noi due è identico a prima.

La stessa domanda



Il problema, ogni giorno

Ogni addebito è una domanda: condivisa, mia, o di Virginia? Sbagliare non rompe niente e non dà errori: ruba soldi in silenzio a me o a lei.

Perché non un foglio di calcolo

Perché swipare tipo Tinder è molto meglio

la proposta non esce da un modello: esce da un conteggio

8 anni

di storico

1.818

spese in archivio

113.496 €

transati

589

movimenti rivisti

Undici funzioni, e un posto solo dove sta il dato

LE FONTI

Fineco
Satispay
Glovo



VERCEL — 11 FUNCTION SU 12

statico + PWA
movements (dispatch)
expenses, balance,
categories, backups,
export/excel
cron 06:00 UTC



SUPABASE

expenses – il mastro
movements – la coda
merchant_rules
ip_bans, enrichment
expenses_balance()
expenses_pivot()

Il tetto di dodici funzioni del piano gratuito ha deciso la forma dell'API.

Tutta la revisione sta in un endpoint solo.

Non è un'eleganza: la dodicesima funzione è tenuta libera per quando servirà.

Tre fonti, tre compromessi diversi

COME ARRIVANO I MOVIMENTI, E PERCHÉ NON MEGLIO DI COSÌ

FINECO · 550	Automatica via open banking by Budgetbakers. Finestra di 35 giorni.
SATISPAY · 27	Caricato Excel
GLOVO · 12	Nessuna API o Excel: si copia/incolla dal web.

Nessuna delle tre arriva come la vorrei.
L'architettura è fatta da quello che concedono.

Perché non sincronizza ogni ora

La PSD2 limita a quattro i risvegli giornalieri non presidiati.

Il secondo innesco

Quando apro la revisione dal telefono sono presente: quello è un accesso presidiato e non ha limiti.

Che cosa succede a un addebito, da quando esiste

IL PERCORSO, SENZA SCORCIATOIE

arriva	sync Fineco, import Satispay
deduplica	per identificativo e per importo
classifica	regole, poi statistiche del negozio
si cataloga	solo se alta confidenza
decido io	destra condivisa, sinistra mia
diventa spesa	subito, non a fine sessione

Lo stato dice dov'è

da_rivedere, deciso, caricato. E «caricato» significa verificato in autonomia.

A oggi

589 movimenti sono passati di qui: 214 sono diventate spese condivise, 375 sono rimaste mie. In coda da rivedere: zero.

Chi decide se una spesa è condivisa

NESSUN MODELLO IN TUTTO IL PERCORSO

```
lib/classifier.js

if TECH_RULES.test(desc)      → no
    trading, canone, bolli, estratto conto

if /pagopa/.test(desc)        → virginia multe
if lunedì, dalle 16           → no palestra

merchant_rules[negozio]
    si 29 / no 1               → si, alta
    si 25 / no 47             → rivedere
```

in tutto il percorso non c'è una sola chiamata a un LLM

Il dominio è fatto di regole

Il lunedì è quasi sempre una spesa con amici della palestra, non con Virginia

E le eccezioni pure

Con l'ora nota quella regola vale solo dalla sera: un pranzo di lunedì non è una cena di lunedì.

Perché nessun modello LLM

La risposta non sta nel significato delle parole: sta in quello che è successo prima misurato attraverso le statistiche.

Quando decide e quando preferisce tacere

La stessa proposta, due trattamenti diversi: uno salta la revisione umana, l'altro no. La differenza è un rapporto misurato.

PROPOSTA AD ALTA CONFIDENZA

SALTA LA REVISIONE

si cataloga da sola, e diventa spesa

119/120

decisioni confermate

PROPOSTA A MEDIA CONFIDENZA

PASSA DALLA REVISIONE

la decido io, una per una

105/153

decisioni confermate

E 316 volte su 589 non propone niente: si astiene più della metà delle volte.

La soglia non è un'impressione: è questo rapporto. Sbaglia una volta su centoventi, quindi non mi chiede il permesso.

Impara soltanto da quello che ho deciso io

CHE COSA AGGIORNA I CONTATORI DI UN NEGOZIO

lo swipe	la mia decisione: questa insegna
l'automatismo	registra la spesa, non insegna
l'import storico	ha seminato le regole, una volta

Se le sue proposte contassero come prove, il sistema si darebbe ragione da solo: confermerebbe per sempre il primo errore.

Quello che sa oggi

337 negozi, 597 decisioni registrate nei contatori. Tredici restano ambigui, e quelli tornano in coda ogni volta.

Storico GLOVO: 25 sì, 47 no

Non imparerà mai, perché la risposta non dipende dal negozio: dipende dall'indirizzo di consegna casa/ufficio, serve una modifica al motore

Il telefono non deve scaricare otto anni di spese

LA STESSA STATISTICA, CALCOLATA IN DUE POSTI DIVERSI

SPESE GREZZE AL CLIENT

463 KB

›

PIVOT CALCOLATE IN SQL

9 KB

Somma e raggruppamento li fa il database con due funzioni SQL, non il browser di chi è in giro con il telefono.

E la trappola sotto, che non dà nessun errore:

PostgREST su supabase tronca a mille righe qualunque select, anche con un limite più alto. Il dato manca ma la risposta sembra completa.

Lo stesso guasto di prima

Non un errore: un'assenza. Il saldo sbagliato sarebbe arrivato con la suite di controllo tutta verde.

Mobile prima di tutto

Quest'app si usa quasi solo da un telefono, spesso dentro un negozio, con una tacca di rete.

GRAZIE



DARIO MARTINELLI