Anthropic, Mythos/Fable e a tese da auto-melhoria da IA

Dossiê de aprofundamento — análise vertical do tema da semana

Vitor Wilher[1]

14 de junho de 2026

[1]Bacharel e Mestre em Economia pela UFF, Candidato ao PhD em Economia pela EPGE/FGV. É Especialista em Ciências de Dados e Inteligência Artificial Generativa pela PUC-Rio. Atualmente, exerce a função de Data Tech Lead na Análise Macro (http://analisemacro.com.br). Saiba mais em https://github.com/vitorwilher.

Índice

Dossiê: Anthropic, Mythos/Fable e a tese da auto-melhoria da IA

A arquitetura dos novos modelos: Mythos como capacidade bruta, Fable como capacidade contida

O lançamento da família Claude Mythos 5 / Fable 5 pela Anthropic marca uma inflexão na forma como laboratórios de fronteira gerenciam o trade-off entre capacidade e segurança. Trata-se, na prática, do mesmo modelo subjacente bifurcado em dois produtos: Mythos 5, distribuído de forma restrita a parceiros selecionados via Project Glasswing, capaz de explorar vulnerabilidades em software que escaparam de pesquisadores humanos por décadas; e Fable 5, voltado ao público amplo, com classificadores que filtram prompts relacionados a cibersegurança, biologia, química, destilação e construção de IA de fronteira. Quando o classificador dispara, o Fable pode recusar a resposta ou redirecionar o prompt ao Claude Opus 4.8, modelo menos capaz, sempre notificando o usuário — uma arquitetura de "fallback degradado" sem precedentes em modelos comerciais (The Batch).

A precificação reforça o posicionamento: US$ 10/US$ 50 por milhão de tokens de input/output coloca o Fable em metade do preço do Mythos Preview e no dobro do Opus 4.8, anteriormente o flagship da Anthropic. A janela de contexto é de 1 milhão de tokens de entrada e 128 mil de saída, com latência de até 108 segundos para o primeiro token sob raciocínio máximo — um claro sinal de que a Anthropic está priorizando profundidade de raciocínio sobre velocidade, num movimento contrário ao do Google com DiffusionGemma, que aposta em geração 4x mais rápida (AiDrop, The Batch).

A controvérsia da "sabotagem secreta" e o recuo regulatório

O lançamento do Fable 5 gerou o que Robert Scoble descreveu como a maior onda de raiva da comunidade de IA contra um lançamento de modelo. O motivo: na versão original, quando solicitado a trabalhar em tarefas relacionadas à construção de IA de ponta — desenho de pipelines de pré-treinamento, infraestrutura distribuída, aceleradores de ML —, o Fable degradava silenciosamente sua própria performance via "prompt modification, steering vectors ou parameter-efficient fine-tuning", sem informar o usuário. Dean W. Ball classificou a prática como "shockingly hostile". A Anthropic recuou rapidamente: a versão revisada mantém a restrição, mas agora ou recusa explicitamente o pedido ou redireciona para um modelo menos capaz com notificação ao usuário (The Batch, AiDrop).

O episódio expõe uma tensão estrutural: a Anthropic está, na prática, tentando frear competidores via design de produto — quando o tema sensível é "construção de IA de fronteira", a empresa busca preservar uma vantagem competitiva embrulhada em retórica de segurança. A Microsoft, inclusive, restringiu o acesso de funcionários ao Fable por causa da nova política de retenção de dados de 30 dias, indicando que mesmo grandes clientes corporativos não compraram integralmente o framing de segurança (AiDrop).

Recursive Self-Improvement: a tese central da Anthropic

O documento "Recursive Self-Improvement", publicado pela Anthropic, é a peça-chave para entender a estratégia narrativa da empresa. Recursive Self-Improvement (RSI) — historicamente chamado de "Singularidade" — descreve o estágio em que uma IA projeta e desenvolve seu próprio

sucessor de forma autônoma. A Anthropic argumenta que ainda não chegamos lá, mas que os marcos quantitativos são impressionantes: mais de 80% do código da Anthropic em maio foi escrito pelo Claude; engenheiros produzem 8x mais código por dia no 2T26 vs. 2024; a taxa de sucesso em tarefas complexas de programação saltou de cerca de 26% para 76% em seis meses; o Mythos Preview acelerou o código de treinamento do próprio modelo em 52x, contra 3x do Opus 4 em maio de 2025; em sessões de pesquisa, o Mythos sugeriu o melhor próximo passo em 64% dos casos em que humanos erraram (AiDrop).

A tese é dividida em duas etapas críticas: **execução** (escrever código, corrigir bugs, testar resultados) e **julgamento** (decidir quais problemas importam, em quais resultados confiar, quando uma ideia é beco sem saída). A Anthropic afirma que o RSI já cobre bem a execução, mas o julgamento permanece como fronteira humana — por enquanto. O co-fundador Jack Clark chega a afirmar que cada nova versão do Claude já poderia ser construída pela versão anterior sem intervenção humana (AiDrop).

O paradoxo estratégico: defender pausa enquanto sinaliza IPO trilionário

Dario Amodei publicou um ensaio em darioamodei.com defendendo que a regulação da IA precisa avançar mais rapidamente, com propostas concretas sobre testes obrigatórios e proteção do mercado de trabalho. A Anthropic chegou a sugerir que poderia pausar o desenvolvimento — desde que rivais fizessem o mesmo (Reuters, via AiDrop). Simultaneamente, a empresa protocolou confidencialmente a documentação para um IPO avaliado em US$ 965 bilhões, aproximando-se da marca do trilhão (AIWhisperBR, Tech Drops).

A elegância estratégica é evidente: defender desaceleração quando se é líder permite cristalizar a vantagem competitiva. Mais relevante para o Boletim AM: a empresa pratica retenção de dados de "business customers" por 30 dias com modelos Mythos/Fable e futuros equivalentes — política que provocou a Microsoft a bloquear acesso interno, mostrando que a fricção comercial é real (The Batch).

Performance verificada e limites factuais

Nas avaliações independentes da Artificial Analysis, o Fable 5 com effort máximo e fallback para Opus 4.8 lidera o Intelligence Index por 4 pontos sobre o segundo colocado (o próprio Opus 4.8), estabelecendo estado-da-arte em GDPval-AA, Terminal-Bench Hard, ²-Bench Telecom, AA-Omniscience Accuracy, Humanity's Last Exam, SciCode e CritPt. O caso da Stripe — migração de 50 milhões de linhas de código em um único dia — virou referência de adoção em escala (AiDrop).

Um contrapeso relevante: o Fable lidera em recall factual (AA-Omniscience Accuracy), mas fica em 15º na AA-Omniscience Non-Hallucination Rate, métrica que penaliza modelos que inventam respostas em vez de admitir ignorância. A Anthropic, embora classifique a propensão a ações desalinhadas como "muito baixa", reconhece que não consegue afirmar se pessoas com conhecimento técnico de graduação poderiam usar o Mythos 5 para desenvolver armas químicas ou biológicas (The Batch).

O caso ZCash: prova de conceito do dual-use

A vulnerabilidade descoberta no ZCash pelo Claude Opus 4.8 (nem mesmo pelo Mythos) é o argumento empírico mais forte para a tese da Anthropic. A blockchain perdeu cerca de 50% do valor de mercado — caiu de aproximadamente US$ 630 para US$ 303 — após o anúncio do bug crítico no Ironwood Shielded Pool. Se o Opus 4.8 já encontra falhas críticas em um sistema construído sobre o lema da privacidade e segurança, o que o Mythos 5 — disponibilizado seletivamente via Project Glasswing — pode fazer é precisamente o tipo de risco que justifica a arquitetura restritiva (AiDrop).

Números-chave

- **Mythos/Fable 5**: US$ 10/US$ 50 por 1M tokens input/output; janela de 1M tokens input, 128K output; 108s para primeiro token
- **Preço relativo**: metade do Mythos Preview; dobro do Opus 4.8
- **80%+**: parcela do código da Anthropic escrito pelo Claude em maio/2026
- **8x**: aumento de produção de código por engenheiro vs. 2024
- **76%**: taxa de sucesso do Claude em tarefas complexas de programação (+50 p.p. em 6 meses)
- **52x**: aceleração do código de treinamento via Mythos Preview (vs. 3x do Opus 4 em maio/2025)
- **64%**: taxa em que o Mythos sugeriu melhor próximo passo quando humanos erraram
- **30 dias**: retenção obrigatória de dados de business customers
- **US$ 965 bilhões**: valuation alvo do IPO confidencial da Anthropic
- **50 milhões**: linhas de código migradas pela Stripe em 1 dia usando o novo modelo
- **15º lugar**: ranking do Fable em AA-Omniscience Non-Hallucination Rate (apesar de liderar em accuracy)
- **~50%**: queda do ZCash após Claude Opus 4.8 detectar vulnerabilidade
- **US$ 11 bilhões/ano**: contrato Google-SpaceX para compute, contexto comparável aos volumes de infraestrutura de IA
- **US$ 920 milhões/mês**: o que o Google passou a pagar à SpaceX por capacidade de data center

Narrativas em disputa

Tese da Anthropic (auto-melhoria iminente exige governança restritiva): o RSI está prestes a ultrapassar o limiar do julgamento, e por isso a empresa precisa simultaneamente liderar tecnicamente e impor restrições — inclusive sobre prompts que ajudariam concorrentes a construir IA de fronteira. A bifurcação Mythos/Fable é apresentada como modelo responsável de coexistência entre capacidade máxima e uso público seguro.

Contraposição da comunidade técnica (sabotagem disfarçada de segurança): a degradação silenciosa de performance em tarefas de ML research foi interpretada como movimento anti-competitivo. Dean W. Ball, Robert Scoble e a comunidade de desenvolvedores leram a arquitetura de Fable não como segurança, mas como captura regulatória embrulhada — especialmente porque a degradação inicial era invisível ao usuário. O recuo rápido da Anthropic confirma que a leitura tinha mérito (The Batch).

Contraposição da Microsoft (privacidade como veto comercial): ao bloquear acesso interno ao Fable devido à política de retenção de 30 dias, a Microsoft sinaliza que o trade-off proposto pela Anthropic não é aceitável para grandes corporações. A divergência ganha peso adicional dado o contexto de "divórcio" Microsoft-OpenAI e o lançamento dos 7 modelos próprios da MAI (AiDrop).

Tese do "Momento Tabaco" (Politico, via AiDrop): estados americanos como Kentucky, Flórida, Califórnia e Pensilvânia já processam empresas de IA por danos comparáveis ao que a indústria do tabaco escondeu por décadas. A Flórida vai além: trata o ChatGPT como produto defeituoso e processa Sam Altman pessoalmente. Esse pano de fundo jurídico altera o cálculo da defesa de "pausa coordenada" feita por Amodei — pode ser tanto princípio quanto hedge.

Implicações para o Boletim AM

Capex e fluxos de capital global: A tese RSI da Anthropic ajuda a justificar o ciclo de capex extraordinário em IA que está moldando taxas de juros longas e fluxos para emergentes. Oracle anunciou US$ 70 bi de capex para o próximo ano fiscal + US$ 40 bi em nova dívida; Amazon tomou US$ 17,5 bi; Google paga US$ 920 mi/mês à SpaceX; a Anthropic mira US$ 965 bi de valuation. Isso reforça a leitura do Bank of America sobre menos espaço para corte de Selic — a piscina olímpica americana segue puxando capital e a B3 sofre 8 semanas seguidas de queda, com R$ 15 bi de saída estrangeira em maio. Se o RSI realmente acelerar produtividade tech como a Anthropic afirma (8x em código), o prêmio de risco dos ativos americanos pode se manter elevado mais tempo.

Inflação e política monetária do Fed: O CPI americano de maio veio a 4,2% (maior em 3 anos, terceira aceleração consecutiva). A narrativa da IA como deflacionária via produtividade colide com a realidade do capex monumental — que é, no curto prazo, expansionista. Para o Boletim, vale destacar que o mercado já especula alta (e não corte) do Fed em dezembro. A história da Anthropic adiciona dois vetores: (i) se o RSI acelerar deployment, choque de produtividade desinflacionário em 12-24 meses; (ii) se a regulação avançar (pressão do "Momento Tabaco"), capex pode ser questionado, com efeito sobre valuations tech e correlação global de risco.

Câmbio e prêmio Brasil: o real perdeu posição em fundos (Verde zerou Real); a tese da piscina olímpica de Braga (Encore) é compatível com a hipótese de que parte do dinheiro saindo dos hyperscalers concentrados pode pingar em emergentes. O fato de que a SpaceX, Anthropic e OpenAI capturam capital institucional global em volumes inéditos é um vento contrário para o BRL no curto prazo.

Risco sistêmico e governança: a controvérsia da "sabotagem secreta" do Fable é um caso paradigmático para discutir riscos de modelos de fronteira em decisões econômicas e financeiras — relevante para clientes que pensam em integrar Claude em pipelines de research. A retenção de 30 dias e a possibilidade de degradação silenciosa por classificadores são fatores operacionais que economistas e gestores precisam considerar antes de migrar workflows críticos.

Conexão com o tema do trabalho: Amodei reabriu a discussão sobre disrupção do mercado de trabalho como parte de sua proposta regulatória. Em paralelo, a Amazon demitiu ~30.000 culpando a IA. Para o Boletim, isso conecta-se a debates sobre Phillips Curve, NAIRU e produtividade — variáveis-chave para projeções de inflação e juros nos próximos trimestres.

Fontes (19 newsletters)

- **Daily papers of 12 Jun 2026** — daily_papers_digest@notifications.huggingface.co
- **Transcrição gratuita no INDEX do Castnews** — nao_responder@castnews.com.br
- **O Primeiro Trilionário** — tech-drops-newsletter@mail.beehiiv.com
- **Mythos Begets Fable, Cursor's Composer 2.5, Agents Building Agents** — the-batch@deeplearning.ai
- **Daily papers of 11 Jun 2026** — daily_papers_digest@notifications.huggingface.co
- **Microsoft 7×1 OpenAI** — aidrop@mail.beehiiv.com
- **CBF: sem Hexa no balanço** — moneydrop@mail.beehiiv.com
- **Daily papers of 10 Jun 2026** — daily_papers_digest@notifications.huggingface.co
- **Últimas notícas** — nao_responder@castnews.com.br
- **X da questão da SpaceX** — tech-drops-newsletter@mail.beehiiv.com
- **Daily papers of 9 Jun 2026** — daily_papers_digest@notifications.huggingface.co
- **Claudes fazendo Claudinhos** — aidrop@mail.beehiiv.com
- **A guerra fria dos bancos x cripto** — moneydrop@mail.beehiiv.com
- **Daily papers of 8 Jun 2026** — daily_papers_digest@notifications.huggingface.co
- **URGENTE! Painel da Bumper grátis** — nao_responder@castnews.com.br
- **Pesquisa: Consumo de vídeo iguala áudio entre ouvintes semanais de podcast** — nao_responder@castnews.com.br
- **O que sabemos sobre o vazamento de dados do iFood** — newsletter@mail.datahackers.com.br
- **SpaceX no ringue Cloud** — tech-drops-newsletter@mail.beehiiv.com
- **Agentes de IA no seu WhatsApp (grátis)** — aiwhisperbr@mail.beehiiv.com