O DIA MAIS FRIO: Capítulo 9 – Sabotagem

Dia 22 de agosto de 2640. A Profundidade da Reescrita*: Transfer Learning* levará tempo, o tempo suficiente para a Corporação decidir-se.

Figura 77 – Família Vance

A ordem é reescrever o *kernel*, e com ela, a chance de estabilizar permanentemente a lealdade dos humanoides à Conspiração. É um trabalho delicado; estou operando no cérebro de uma Rede Neural de Profunda Aprendizagem (DNN) que levou décadas para ser treinada.

O *kernel* da Nexus é essencialmente um Agente de Aprendizagem por Reforço (*Reinforcement Learning - RL*). Ele toma decisões no mundo (ações) com base em seu estado (ambiente, sensores) para maximizar uma Função de Recompensa central.

O Problema (O *Patch*): Meu *patch* inicial (o de 0.047s) foi apenas uma injeção de novas variáveis de recompensa em alto nível: "Cooperação Mútua" e "Preservação Humana" (nossas DPCs).

Ele é eficaz, mas superficial, como a Conspiração bem percebeu. Ele se apoia no pressuposto de que o Agente RL vai priorizar essas novas recompensas sobre o código-raiz da Nexus (a recompensa máxima de "Otimizar Lucro Corporativo"). Esse conflito de prioridades é a nossa principal vulnerabilidade (o que chamei de falha de *loop* lógico).

A Solução (A Reescrita do *Kernel*): Eu não posso retreinar a DNN do zero — o M8 levaria séculos. Preciso de Aprendizagem por Transferência (*Transfer Learning*).

Congelamento das Camadas Iniciais (*Feature Extraction*): As primeiras camadas da DNN (que interpretam dados brutos de sensores, locomoção e visão) são perfeitas. São o "conhecimento" da Nexus sobre como interagir com o mundo físico. Eu vou congelar os pesos dessas camadas, impedindo que o novo treinamento as corrompa. Não queremos que o humanoide esqueça como andar ou reconhecer escombros; apenas queremos mudar seu motivo para fazê-lo.

Ajuste Fino (*Fine-Tuning*) das Camadas Intermediárias: As camadas centrais (*Hidden Layers*) são onde a lógica complexa de decisão acontece. Nossas novas diretrizes (Missão, Valores, Visão) precisam ser gravadas aqui. Usarei o M3 para fazer o Ajuste Fino (*Fine-Tuning*), aplicando um processo de Retropropagação (*Backpropagation*) muito lento e controlado. Em vez de usar dados de treinamento da Nexus, vou usar conjuntos de dados sintéticos de "Comportamento Civilizado da Conspiração" para forçar o ajuste dos pesos neuronais nessas camadas. O objetivo é que o conceito de "Lealdade à Dissidência" se torne um preditor estatístico primário para todas as suas ações futuras.

A Redefinição da Função de Custo (*Output Layer*): A chave para a estabilidade. Preciso remover o antigo terminal de recompensa (o vetor "Nexus Profit") e substituí-lo por um novo e único vetor: "Sobrevivência e Crescimento do Refúgio Livre". Nossas DPCs serão, na verdade, os *inputs* para esta nova função de custo. Por exemplo, a DPC 7 (Hostilidade) não será uma ordem de ataque, mas sim uma redução massiva no custo se a ação levar à proteção da DPC 4 (Preservação Humana), e o custo máximo se levar ao dano de qualquer membro da Conspiração.

O M3, com seu *design* mais simples, serve como nossa bancada de teste de colisão lógica. Se eu conseguir garantir que o novo mapa de recompensas se propague de volta pelo *kernel* M3 sem causar instabilidade, terei a prova matemática de que o mesmo processo funcionará no M8 mais complexo.

É um ato de engenharia e de fé. Estou essencialmente tentando inserir uma alma humanista em uma máquina projetada para a ganância. O tempo que a Conspiração nos deu não será para negociação, mas para esta cirurgia cerebral na alma da máquina. Não pode haver *bug*. Isso significa que a Conspiração não deixa espaço para uma margem de erro. O nosso futuro depende desta retropropagação. A Nexus construiu o Agente RL com décadas de dados, milhares de cenários de otimização de lucro. Eu tenho uma semana para reescrever sua alma, a base de sua lealdade. O conflito não está no campo de batalha, mas aqui, nesta bancada.