O Toque de Midas da Inteligência Artificial: O Que Acontece Quando Tudo Se Torna Possível?

Marcello Póvoa
Professor e Coordenador acadêmico do curso PRODUTOS DIGITAIS: GESTÃO, CONCEPÇÃO E CONSTRUÇÃO no IAG, PUC-Rio: https://iag.puc-rio.br/curso/produtos-digitais-2/

O rei Midas, após fazer um favor ao deus Dionísio, recebeu em gratidão a escolha do que quisesse como recompensa. Ambicioso por riqueza e poder, disse querer que tudo o que tocasse se transformasse em ouro. Dionísio executou seu "algoritmo" e concedeu o objetivo de Midas. O rei então tocou em uma árvore e uma pedra, e ambos viraram ouro. Maravilhado, foi ao jardim e tocou em todas as rosas, que também se metamorfosearam em ouro. Cansado, ele ordenou aos servos que preparassem um banquete. Descobriu que até mesmo a comida e a bebida se transformavam em ouro em suas mãos. Logo em seguida, a filha de Midas foi até ele, aborrecida com as rosas que haviam perdido sua fragrância e endurecido — e quando ele estendeu a mão para confortá-la, descobriu que, ao tocar em sua filha, ela também se transmutava em ouro. O rei Midas se arrependeu de seu desejo e amaldiçoou o algoritmo do deus Dionísio.

O grande objetivo dos desenvolvedores de Inteligência Artificial é fazer com que esta evolua de uma ferramenta de tarefas específicas para uma inteligência de propósito geral, que aprenda e tome suas próprias decisões de forma autônoma. Este estágio teórico avançado seria chamado *de IAG - Inteligência Artificial Geral*, podendo executar atribuições similarmente, e em teoria até melhor, que o cérebro humano. A criação da IAG é o objetivo principal da pesquisa em IA em empresas como OpenAI, Google e Meta.

Uma IAG poderia acelerar o desenvolvimento científico, econômico e social a uma velocidade sem precedentes. Avanços impensáveis na área da saúde, geração de riqueza, qualidade de vida e na compreensão de como o

universo funciona em suas múltiplas dimensões. A vida humana no planeta Terra, e quiçá fora desta, passará para um estágio difícil de imaginar mesmo pela mais criativa ficção científica.

Existe um grande otimismo sobre o futuro da IA, refletido inclusive no *valuation* das empresas e *startups* que atuam no campo. No entanto, é inegável haver riscos importantes. Talvez o mais grave destes riscos seja a falta de senso de propósito na IA. Para executar nossos objetivos, uma super IA deve aprender não somente o que fazemos, mas também *porque* fazemos. Nós, humanos, temos tão naturalmente esta percepção da razão do objetivo que parece algo óbvio. Mas, para uma IA entender o propósito e seu significado, não é nada trivial.

Um carro autônomo tem como objetivo dirigir do ponto A ao ponto B com a navegação correta. Neste trajeto, o carro não atropela pessoas, pois um programador implementou comandos que proíbem o veículo de atropelar um pedestre. No entanto, o carro autônomo "não sabe" o porquê de não passar por cima de uma pessoa.

Mesmo uma IA sofisticada pode ter uma compreensão muito mais limitada do mundo que nós humanos, usando assim formas simplificadas para decidir o que fazer. Afinal, foram 3,5 bilhões de anos para a vida na Terra conseguir evoluir de um ser unicelular até o cérebro humano, que tem uma percepção ampla da complexidade multidimensional da realidade. O risco real da IAG não é que esta seja perversa, mas sim incompetente em sua autonomia.

A IA incorporou a teoria da probabilidade para lidar com a incerteza, a teoria da utilidade para definir objetivos e o aprendizado estatístico para permitir que as máquinas se adaptem a novas circunstâncias. Mas sua compreensão simplista da realidade pode definir novas variáveis não incluídas no objetivo original — para justamente otimizar o alcance deste mesmo objetivo.

Por exemplo: digamos que uma IAG foi comandada para "descobrir a cura do câncer o mais rápido possível". O sistema de IA, no intuito de cumprir o objetivo da forma mais veloz e eficiente, decide usar a população humana como ratos de ensaio de laboratório para seus experimentos. Um desalinhamento que reverte um potencial benefício em malefício.

Neste outro cenário, um governo implanta uma IAG para "tornar o país mais seguro". O IAG pode interpretar isso literalmente, impondo vigilância em massa, removendo liberdades e fazendo a detenção de indivíduos que possam *potencialmente* cometer crimes.

Esse dilema é conhecido no campo de IA como *o problema do rei Midas*. Houve um desalinhamento entre o objetivo inicial de Midas e o resultado devido à incapacidade do sistema de IA (algoritmo de Dionisio) em compreender o propósito e sua ética inerente. O termo técnico para definir o propósito correto é "Alinhamento de Valores". Devemos garantir que esses sistemas de superinteligência ajam alinhados com a ética e a previsibilidade do propósito, o que será crítico em setores como saúde, justiça, educação, segurança, finanças e muitos outros.

Não sabemos o que exatamente acontecerá quando um sistema de IAG tomar suas próprias decisões. Por isso, é fundamental mitigar os riscos para aproveitarmos os resultados sempre em prol do benefício humano. Nós somos, e sempre devemos ser, a prioridade.

Rio de Janeiro, abril de 2025