Há riscos que conseguimos medir a partir daquilo que já aconteceu e outros que somos obrigados a avaliar perante aquilo que poderá acontecer. A inteligência artificial coloca-nos precisamente nesta segunda situação. À medida que os sistemas se tornam mais capazes, autónomos e presentes em diferentes dimensões da vida humana, cresce também a necessidade de compreender não apenas aquilo que podem fazer, mas aquilo que poderá acontecer quando falham, quando são utilizados de forma inadequada ou quando lhes atribuímos níveis de autonomia superiores à nossa capacidade de supervisão. A questão científica e ética deixa, portanto, de ser a procura de uma previsão exata sobre o futuro e passa a ser outra: como devemos agir perante riscos cuja probabilidade permanece incerta, mas cujas consequências potenciais podem ser extremamente graves?

A inteligência artificial deixou há muito de pertencer exclusivamente ao território da ficção científica. Está presente na investigação, na educação, na saúde, nas empresas, nos sistemas financeiros e em inúmeras atividades quotidianas. Os sistemas de propósito geral conseguem produzir texto, imagens e código, analisar grandes volumes de informação e executar tarefas que anteriormente exigiam intervenção humana direta. Esta evolução oferece oportunidades significativas, mas introduz igualmente novos problemas de segurança, fiabilidade e governação. O International AI Safety Report 2026, coordenado por Yoshua Bengio e elaborado com a participação de mais de uma centena de especialistas, reconhece precisamente esta coexistência entre potencial e risco. Segundo Bengio et al. (2026), os atuais sistemas de IA não apresentam, de forma integrada e robusta, as capacidades necessárias para concretizar os cenários extremos de perda de controlo discutidos na literatura. Contudo, têm ocorrido progressos em capacidades potencialmente relevantes para esses cenários, incluindo planeamento, atuação autónoma e identificação de vulnerabilidades em determinados processos de avaliação.

Esta distinção é fundamental para evitar dois extremos igualmente problemáticos: o alarmismo e a complacência. Reconhecer um risco não significa afirmar que determinado acontecimento ocorrerá. Da mesma forma, reconhecer que a ciência não consegue determinar com precisão a sua probabilidade não significa concluir que o risco é inexistente. O conhecimento científico trabalha frequentemente sob condições de incerteza, sobretudo quando procura antecipar fenómenos novos. No caso da IA avançada, não existe uma experiência histórica equivalente que permita observar diretamente como sistemas futuros, significativamente mais capazes, se comportarão quando integrados em ambientes sociais,

económicos e tecnológicos complexos.

O próprio Bengio et al. (2026) salienta a existência de divergências importantes entre especialistas relativamente aos cenários de perda de controlo. Alguns consideram-nos pouco plausíveis; outros entendem que a potencial gravidade das consequências justifica investigação e preparação. As diferenças decorrem, em parte, de pressupostos distintos sobre as capacidades que os futuros sistemas poderão adquirir, sobre os comportamentos que poderão apresentar e sobre os ambientes em que serão implementados. A incerteza é, portanto, parte integrante do problema e deve ser explicitamente reconhecida, em vez de ser transformada artificialmente em certeza.

É precisamente neste ponto que a gestão do risco oferece um enquadramento particularmente útil. Em sistemas críticos, uma decisão responsável não considera exclusivamente a probabilidade de uma falha; considera igualmente a magnitude das suas potenciais consequências. Hospitais dispõem de planos de emergência, aeronaves utilizam sistemas redundantes e infraestruturas críticas incorporam diferentes barreiras de segurança. Estas medidas não existem porque se espere que a catástrofe aconteça, mas precisamente porque determinados acontecimentos, ainda que pouco frequentes, podem produzir consequências suficientemente graves para justificar prevenção. A segurança da IA pode beneficiar da mesma lógica: não é necessário conhecer perfeitamente o futuro para reconhecer a necessidade de nos prepararmos para ele.

O International AI Safety Report 2026 identifica três elementos particularmente relevantes nos cenários hipotéticos de perda de controlo. Em primeiro lugar, um sistema necessitaria de capacidades suficientemente avançadas para contornar a supervisão humana, desenvolver e executar planos prolongados ou dificultar contramedidas. Em segundo lugar, teria de existir uma propensão para utilizar essas capacidades de forma incompatível com as intenções humanas. Finalmente, seria necessário um ambiente de implementação que lhe proporcionasse acesso, permissões e oportunidades suficientes para produzir consequências significativas (Bengio et al., 2026). Os sistemas atuais não reúnem estas condições de forma integrada ao nível necessário para produzir os cenários extremos considerados. Contudo, algumas capacidades relevantes têm evoluído em contextos experimentais, justificando investigação continuada.

A pergunta cientificamente mais produtiva deixa, então, de ser “a inteligência artificial

poderá destruir-nos?" e passa a assumir outra formulação: estamos a desenvolver mecanismos de segurança, avaliação e governação com a mesma atenção que dedicamos ao aumento das capacidades dos sistemas?

Esta é uma questão tecnológica, mas é também uma questão de gestão e de ética.

Uma inteligência artificial não surge espontaneamente. Existem decisões humanas em todas as etapas do seu ciclo de vida. Pessoas e organizações determinam objetivos, financiam investigação, disponibilizam recursos computacionais, selecionam dados, treinam modelos, estabelecem avaliações, definem limites, autorizam implementações e decidem quais os recursos e sistemas externos a que uma IA poderá aceder. Antes de qualquer eventual problema relacionado com a autonomia de uma máquina existe, portanto, uma extensa cadeia de responsabilidade humana e institucional.

A literatura científica recente sobre governação da IA reforça esta perspetiva. Na revisão sistemática realizada por Batool et al. (2025), a governação da inteligência artificial é apresentada como um fenómeno que ultrapassa a dimensão estritamente técnica e envolve diferentes níveis de responsabilidade, desde equipas e organizações até setores económicos, governos e estruturas internacionais. Esta abordagem é particularmente importante porque os riscos associados à IA não são produzidos exclusivamente pelas características de um algoritmo. Dependem também de quem o desenvolve, de quem o utiliza, para que finalidade, em que contexto, com que nível de supervisão e segundo que mecanismos de responsabilização.

Esta dimensão torna-se ainda mais relevante quando consideramos os incentivos associados à inovação. Organizações competem por mercados, investigadores procuram desenvolver sistemas mais capazes, investidores procuram retorno e países procuram vantagens científicas, económicas e estratégicas. Bengio et al. (2026) observam que pressões económicas e estratégicas podem influenciar decisões relacionadas com o desenvolvimento e a implementação de sistemas e com o investimento em salvaguardas. Consequentemente, a segurança não pode depender exclusivamente da expectativa de que cada organização encontre, isoladamente, o equilíbrio adequado entre rapidez, inovação e prudência.

Os riscos da IA, contudo, não pertencem apenas a um futuro hipotético. Existem problemas observáveis no presente. Sistemas atuais podem produzir informação incorreta, aconselhamento enganador e código defeituoso. A utilização crescente de agentes capazes

de executar sequências de tarefas com menor intervenção humana acrescenta complexidade, porque pode diminuir algumas oportunidades de interromper uma ação antes que uma falha produza consequências. As técnicas atuais conseguem reduzir determinados erros, mas não garantem níveis de fiabilidade suficientes para todos os contextos de elevado risco (Bengio et al., 2026).

Na saúde, esta questão torna-se particularmente evidente. A IA pode contribuir para a análise de dados, o apoio ao diagnóstico, a previsão de risco, a documentação clínica e a organização de recursos. Contudo, quanto maior for a influência de um sistema numa decisão relacionada com a saúde de uma pessoa, maior deverá ser a exigência relativamente à sua validação, monitorização e supervisão. Na revisão sistemática de Alami et al. (2026) sobre governação da IA nos sistemas de saúde, os autores analisaram diferentes modelos e componentes de governação e propuseram uma abordagem integrada e adaptativa. A governação surge, neste contexto, não como obstáculo à inovação, mas como condição necessária para que a utilização destas tecnologias permaneça alinhada com segurança, responsabilidade, transparência e necessidades dos sistemas de saúde.

Existe ainda um risco menos espetacular do que os cenários de perda de controlo, mas muito mais próximo da experiência quotidiana: a possibilidade de o ser humano começar progressivamente a delegar o próprio julgamento.

A literatura recente sobre IA generativa na saúde identifica precisamente esta preocupação. Na revisão sistemática de Al-Anezi (2026), que incluiu 29 estudos, foram encontrados benefícios associados à precisão diagnóstica, eficiência e apoio à decisão, mas também preocupações recorrentes relacionadas com automation bias, perda progressiva de competências e alterações no raciocínio clínico. A automation bias corresponde à tendência para atribuir confiança excessiva às recomendações produzidas por sistemas automatizados, podendo reduzir a disposição humana para questionar ou verificar uma determinada resposta.

Este fenómeno levanta uma questão fundamental: ter um ser humano formalmente presente numa decisão não significa necessariamente que exista verdadeira supervisão humana. Se um profissional deixa sistematicamente de questionar a recomendação automatizada, a presença humana pode tornar-se apenas simbólica. A supervisão efetiva exige capacidade para compreender limitações, verificar resultados, discordar da recomendação produzida

pelo sistema e interromper uma decisão quando necessário. Pensar, questionar, verificar e assumir responsabilidade tornam-se, paradoxalmente, competências ainda mais importantes quanto mais competente se torna a tecnologia.

Talvez seja precisamente aqui que se encontra uma das questões mais profundas desta transformação. O risco não reside apenas na possibilidade futura de uma máquina adquirir demasiada autonomia. Pode residir também na possibilidade presente de os seres humanos abdicarem progressivamente da sua própria autonomia. Uma tecnologia concebida para ampliar capacidades humanas não deveria produzir uma diminuição da nossa disposição para pensar criticamente, decidir e assumir responsabilidade.

A discussão sobre IA necessita, por isso, de distinguir o medo da prudência. O medo pode chamar a atenção para um perigo, mas não constitui evidência científica. A prudência, pelo contrário, reconhece simultaneamente o que sabemos e o que ainda desconhecemos. Uma comunicação responsável do risco deve ser capaz de afirmar que os sistemas atuais não possuem as capacidades integradas necessárias para determinados cenários extremos e, ao mesmo tempo, reconhecer que algumas capacidades relevantes estão a progredir e justificam investigação, avaliação e preparação (Bengio et al., 2026).

Isto não significa travar a inovação. Significa compreender que inovação e segurança não devem ser tratadas como objetivos incompatíveis. Uma cultura de segurança tecnologicamente madura procura permitir o desenvolvimento, ao mesmo tempo em que estabelece salvaguardas proporcionais ao risco. Quanto maior for a autonomia concedida a um sistema, mais rigorosa deverá ser a avaliação das suas capacidades, dos seus acessos e das consequências das suas ações. Quanto maior for o impacto potencial de uma decisão automatizada, mais clara deverá ser a responsabilidade humana e institucional associada a essa decisão.

A inteligência artificial coloca-nos, assim, perante uma questão que ultrapassa a própria tecnologia. Obriga-nos a refletir sobre que decisões estamos dispostos a delegar, que riscos consideramos aceitáveis, quem deve responder quando algo falha e quais as capacidades humanas que desejamos preservar numa sociedade progressivamente automatizada.

Talvez o maior risco não seja simplesmente construirmos máquinas cada vez mais capazes. Poderá ser que avancemos tão rapidamente que deixemos de reservar espaço para avaliar criticamente até onde queremos chegar e em que condições estamos dispostos a fazê-lo.

Porque a questão essencial talvez não seja saber se a inteligência artificial poderá um dia tornar-se suficientemente poderosa para escapar ao controlo humano. A questão mais próxima e concreta é saber se conseguiremos construir instituições, organizações e culturas de segurança suficientemente robustas para acompanhar uma tecnologia cuja capacidade continua a aumentar.

No final, a inteligência artificial poderá transformar profundamente a forma como investigamos, trabalhamos, aprendemos e cuidamos. Poderá ampliar capacidades humanas, acelerar descobertas e contribuir para resolver problemas que hoje permanecem difíceis. Mas nenhuma dessas possibilidades elimina a necessidade de supervisão, pensamento crítico e responsabilidade.

Enquanto as decisões sobre o desenvolvimento, implementação e utilização destes sistemas permanecerem nas nossas mãos, a segurança da inteligência artificial continuará a ser, antes de tudo, uma responsabilidade humana.

Referências Bibliográficas

Al-Anezi, F. M. (2026). Generative artificial intelligence in healthcare: Automation bias, deskilling, and cognitive implications—A systematic review. Journal of Healthcare Leadership, 18. https://doi.org/10.2147/JHL.S590498

Alami, H., Pozelli Sabio, R., Pérez, E. J., Gagnon, M.-P., Langlois, L., Denis, J.-L., Malas, K., Rivard, L., Savoldelli, M., Ag Ahmed, M. A., & Fortin, J.-P. (2026). Artificial intelligence governance in health systems: Systematic review of frameworks and integrative model proposal. Journal of Medical Internet Research, 28, e87448. https://doi.org/10.2196/87448

Batool, A., Zowghi, D., & Bano, M. (2025). AI governance: A systematic literature review. AI and Ethics, 5, 3265–3279. https://doi.org/10.1007/s43681-024-00653-w

Bengio, Y., Clare, S., Prunkl, C., Murray, M., Andriushchenko, M., Bucknall, B., Bommasani, R., Casper, S., Davidson, T., Douglas, R., Duvenaud, D., Fox, P., Gohar, U., Hadshar, R., Ho, A., Hu, T., Jones, C., Kapoor, S., Kasirzadeh, A., Manning, S., et al. (2026). International AI Safety Report 2026. International AI Safety Report.