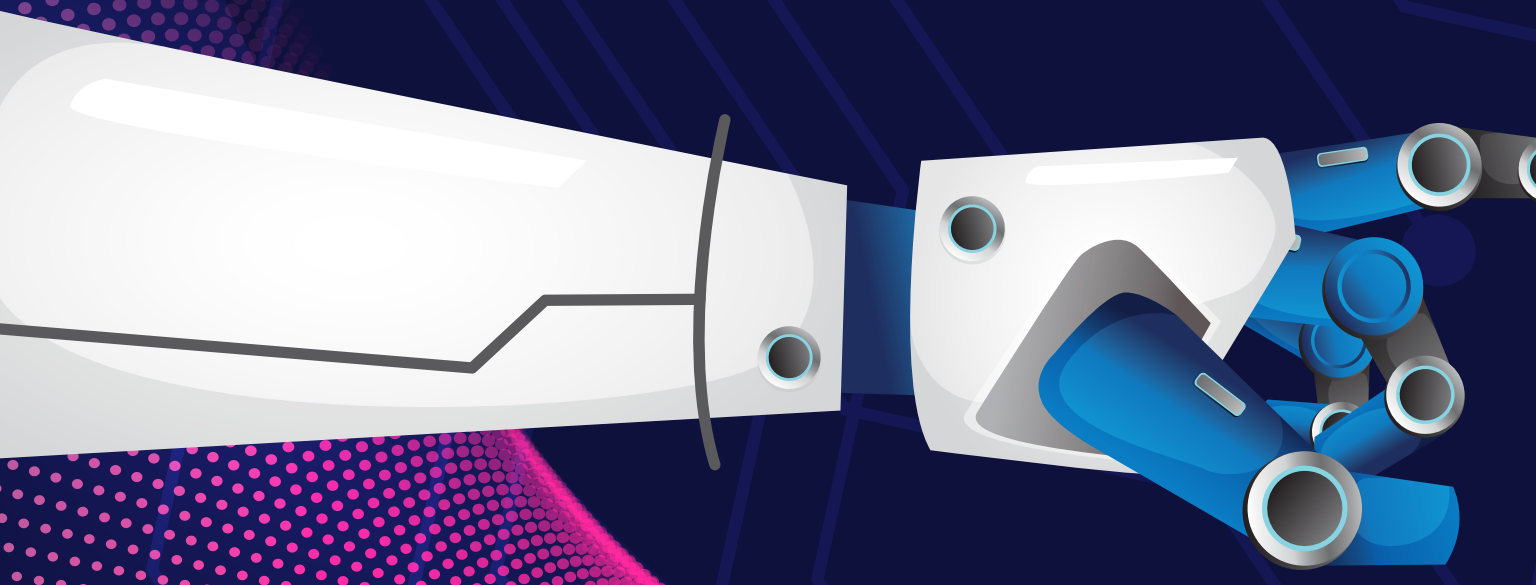
Desenvolvendo a Prontidão para Confiar no Uso da Inteligência Artificial dentro das Equipes de Combate

Capl (Maj) Marlon W. Brown, Exército dos EUA

Estamos no início de uma rápida integração da inteligência artificial (IA) nas operações militares. A *National Security Strategy of the United States* ("Estratégia Nacional de Segurança dos Estados Unidos da América [EUA]") indica que o avanço acelerado no campo da IA, entre outras novas tecnologias, é de extrema importância para a segurança nacional[1]. O *Summary of the 2018 National Defense Strategy of the United States of America* ("Resumo da Estratégia Nacional de Defesa de 2018 dos EUA") ecoa essa preocupação e aborda a necessidade de "investir amplamente na aplicação militar da autonomia, inteligência artificial e aprendizado



de máquina, incluindo a célere aplicação de inovações comerciais, para obter vantagens militares competitivas" como parte da modernização de capacidades-chave, com o intuito de desenvolver uma força mais letal[2].

O Centro Conjunto de Inteligência Artificial está encarregado de executar o recém-elaborado *Summary of the 2018 Department of Defense Artificial Intelligence Strategy* ("Resumo da Estratégia de Inteligência Artificial de 2018 do Departamento de Defesa"). A estratégia inclui a colaboração dos meios de defesa com parceiros acadêmicos e comerciais no desenvolvimento e implementação da nova tecnologia[3]. Um componente dessa abordagem de modernização é a Agência de Projetos de Pesquisa Avançada do Departamento de Defesa dos EUA (*Defense Advanced Research Projects Agency — DARPA*) para a qual o presidente solicitou um orçamento de US$ 3,556 bilhões para o ano fiscal de 2020. O projeto, intitulado "Artificial Intelligence and Human-Machine Symbiosis" ("Inteligência Artificial e a Simbiose Ser Humano-Máquina", em tradução livre), tem um custo previsto de mais de US$ 161 milhões em 2020, um aumento de 233% relação ao orçamento de 2018[4].

Atualmente, a integração da IA é limitada, não tendo, ainda, alterado o combate de maneira significativa, especialmente no nível tático. Os seres humanos ainda detêm o controle total. Como os líderes civis e militares são cautelosos quanto a atribuir à IA qualquer análise ou processo decisório que possa afetar diretamente a vida humana, muitos preveem que essa situação continuará sendo a regra. Entretanto, é provável que esse tipo de parceria entre ser humano e tecnologia mude, porque os adversários desafiarão os EUA com um forte emprego de IA. Independentemente de quantos especialistas de peso em ciência e tecnologia proponham a proibição de armas autônomas ou da qualidade dos possíveis argumentos contra o desenvolvimento de IA, o "gênio da inovação está fora da lâmpada: não há como empurrá-lo de volta"[5]. Os adversários estão investindo fortemente na tecnologia, e os EUA também.

Como as futuras guerras serão caracterizadas pelo emprego de sistemas de IA em rápida evolução, a força militar deve estar pronta para aceitar essa nova tecnologia. A prontidão não se refere apenas a desenvolver e pôr os sistemas mais adequados de IA em serviço. Ela incluirá soluções para questões éticas e morais, como "Os soldados estarão dispostos a combater ao lado de robôs?"[6] Ao responder a esse tipo de pergunta, é preciso considerar a capacidade dos combatentes humanos para confiar nos sistemas artificiais de sua equipe. Tanto a análise do nosso atual conceito doutrinário de confiança dentro de equipes coesas quanto

Imagem cedida *por Army AL&T Magazine*.

a avaliação dos fatores que podem levar a uma decisão individual de confiar possibilitarão que nossos soldados confiem nos sistemas de IA prestes a serem integrados nas equipes de combate.

O que é a IA?

Antes de considerar a questão da confiança na IA, é importante entender o caráter diversificado dessa tecnologia. A tecnologia de IA não é estática, e os rápidos avanços continuam a alterar os requisitos para entendê-la e determinar como a questão da confiança nesses sistemas deve ser tratada. É possível encontrar vários termos para diferenciar entre os tipos e exemplos de IA com uma rápida busca na internet. Um meio útil de classificação, utilizado neste artigo, distingue entre a inteligência artificial estreita ou fraca (IA Estreita, ou *artificial narrow intelligence — ANI*) e a inteligência artificial geral ou forte (IA Geral, ou *artificial generalized intelligence — AGI*). Todos os atuais sistemas de IA operam no campo da IA Estreita, em que o sistema se concentra apenas em tarefas limitadas. O recurso *Siri*, da Apple, um dos sistemas de IA mais conhecidos, só é capaz de realizar um conjunto limitado de tarefas relacionadas aos produtos da empresa. Os sistemas de IA Estreita só podem executar as ações para as quais foram projetados.

A IA Geral, por outro lado, representa o futuro da IA, possibilitando que as máquinas tenham *intenção* e *autoconsciência*. Os sistemas de IA Geral serão como os seres humanos: generalistas e capazes de aplicar as informações aprendidas a uma ampla gama de tarefas e experiências. Com frequência, empregam-se termos filosóficos nos debates sobre IA Geral. Além da intenção e da autoconsciência, termos como *senciência* (capacidade de sentir) e *agência* (poder individual de agir) são palavras comumente encontradas para descrever os sistemas classificados como IA Geral. Em suma, a IA Geral será como o ser humano em termos de emoções e pensamento de nível superior. Personagens fictícios como o Exterminador do Futuro, Wall-e e o Data, da série *Jornada nas Estrelas: A Próxima Geração*, são todos sistemas de IA Geral. Embora muitos sistemas fictícios de IA Geral tenham uma forma humanoide, os sistemas de IA Estreita em desenvolvimento e os futuros sistemas de IA Geral poderão ter componentes robóticos ou projeções audiovisuais ou poderão, ainda, existir no ciberespaço, sem interfaces com aparência humana. A

Página anterior: Composição de Arin Burgess, *Military Review*. Imagens originais cedidas por Harryarts, ddraw e Freepik via https://www.freepik.com/.

confiança na IA Estreita e a confiança na IA Geral terão naturezas diferentes, com base nas definições e experiências do conceito dentro das Forças Armadas[7].

A Doutrina sobre Confiança no Âmbito das Equipes Militares

A doutrina do Exército dos EUA reconhece a importância da confiança no âmbito das equipes militares. A confiança mútua é fundamental para a prática do comando de missão. "A confiança é conquistada ou perdida, mais frequentemente, com ações cotidianas do que com gestos grandiosos ou ocasionais. Advém de experiências e treinamentos bem-sucedidos em comum, sendo normalmente obtida em consequência das operações, mas também pode ser desenvolvida intencionalmente pelo comandante"[8]. A força considera a confiança entre os soldados como "confiança no caráter, competência e compromisso dos profissionais militares para defender e viver de acordo com a Ética do Exército"[9]. É difícil exagerar o nível geral de confiança necessário para desenvolver uma equipe de combate eficaz.

A guerra é um empreendimento humano, mas a integração da IA complica as interpretações históricas sobre a natureza da guerra ao ameaçar substituir pelo menos parte dos integrantes humanos das equipes militares por *hardware* e *software*. Mesmo que a natureza da guerra acabe não sendo afetada pela IA (o que é improvável), prevê-se que o caráter da guerra seja inteiramente afetado por sua integração total. O inventor e escritor Amir Husain sugere que uma das mudanças mais significativas no caráter da guerra em decorrência das crescentes capacidades de IA é a velocidade do combate no nível tático[10]. O que acontecerá quando a mente humana e os sistemas de decisão não conseguirem mais acompanhar as ações das máquinas autônomas do inimigo? Embora as decisões sobre ir à guerra e sobre como conduzir uma operação possam conceder tempo e espaço para a reflexão e análise humana, as unidades táticas podem descobrir que é existencialmente necessário depender da IA para tomar e executar decisões letais no campo de batalha. Em tal cenário, o sistema de IA seria, claramente, um integrante de uma equipe de combate coesa que requer confiança. Portanto, faz-se necessário abordar a questão da confiança entre homem e máquina.

Essa mudança de foco para considerar a confiança em relação a atores não humanos não parece algo tão estranho quando nos damos conta de que ela já existe nas operações militares. O melhor exemplo moderno de confiança mútua entre seres humanos e atores não humanos talvez seja o relacionamento entre os cães de guerra e seus condutores. Relacionamentos muito próximos são forjados entre eles — mais ainda do que os da maioria de donos com seus animais de estimação. O que faz com que a unidade de cães de guerra seja diferente é o nível de confiança que os condutores desenvolvem em relação aos seus animais. Confiam que os cães de guerra não só cumpram as tarefas rotineiras para as quais foram treinados, mas também que protejam seus parceiros humanos em situações de perigo, incluindo o risco de morte.

A confiança que um ser humano pode ter na IA Estreita, que não tem caráter ou compromisso, restringe-se apenas à competência do sistema. A IA Estreita deve demonstrar competência em uma grande variedade de responsabilidades, como, por exemplo, identificar, corretamente, ameaças a meios críticos e formas de mitigá-las. Provavelmente, também acertará na identificação dos atores inimigos no campo de batalha. Além disso, ela pode ser capaz de reconhecer sintomas de depressão entre os integrantes da equipe e recomendar um tratamento.

A confiança na IA Estreita está mais próxima do tipo de confiança que os combatentes podem ter em um determinado sistema de armas ou ferramenta de planejamento do que na confiança que depositam uns nos outros. As ferramentas, sejam feitas de aço ou algoritmos, não devem ser tratadas como verdadeiros "membros" de uma equipe, mesmo quando surge um vínculo afetivo. O grau de apego a um sistema de IA Estreita não muda a natureza do sistema. No filme *Náufrago*, o personagem de Tom Hanks claramente sentiu afeto por uma bola de vôlei, à qual ele carinhosamente deu o nome de "Wilson". Ele talvez tenha até sentido "confiança" em Wilson, contando-lhe seus pensamentos íntimos. Independentemente do grau de apego, Wilson não passava de um objeto de couro e borracha. Era uma ferramenta para manter a sanidade do

O Capl (Maj) Marlon W. Brown, do Exército dos EUA, serve na 2ª Brigada de Combate Blindada, 1ª Divisão de Infantaria, Forte Riley, Estado do Kansas. Concluiu o bacharelado pela East Central University e o mestrado em Divindade pelo Southwestern Baptist Theological Seminary. Serviu como capelão em unidades operacionais de aviação, sustentação (logística), artilharia de campanha e operações psicológicas, além de missões anteriores como oficial de infantaria e de inteligência militar.

náufrago. Embora a IA Estreita seja capaz de agir de forma autônoma, autonomia não equivale a agência. Os combatentes humanos devem ter o cuidado de distinguir entre sua confiança em um sistema de IA Estreita e sua confiança nos integrantes humanos e futuros sistemas de IA Geral.

Entretanto, com a IA Geral, será diferente. Ela terá uma forma de "pessoalidade" que possibilitará tratá-la como um membro de confiança das equipes militares. Atribuir-lhe uma forma de "pessoalidade" não representa, de modo algum, uma tentativa de estabelecer se uma máquina senciente é uma forma de vida ou se ela merece proteções legais como tal. Essas questões éticas devem receber a devida atenção em outros fóruns. Considerar a IA Geral como uma forma de "pessoalidade" é não apenas reconhecer que ela pode ter competência como qualquer IA Estreita, mas também caráter e compromisso. Será capaz de definir e cumprir tarefas diferentes daquelas determinadas pelo comandante ou acordadas pela equipe. Algumas tarefas não serão relacionadas à missão militar. A IA Geral terá objetivos "pessoais" e buscará cumpri-los. Isso pode ser entendido como criatividade. Uma importante parte da capacidade da IA Geral para atuar criativamente e com o caráter valorizado pelos militares será sua capacidade para agir contra seus próprios objetivos, especialmente aqueles relacionados à autopreservação.

Entendendo a Decisão de Confiar na IA

Considerando que a confiança — e, possivelmente, confiança mútua — nos sistemas de IA como parte de uma equipe coesa seja necessária, como os integrantes de equipes de combate podem desenvolver a prontidão individual para confiar em máquinas? Robert F. Hurley criou um modelo que possibilita compreender o fenômeno da confiança e como ela pode ser estabelecida[11]. Seu "Modelo de Decisão de Confiar" (*Decision to Trust Model — DTM*) examina a questão da confiança do ponto de vista tanto da parte que confia (*trustor*) quanto da parte que recebe a confiança (*trustee*). Embora seja de maior utilidade para os relacionamentos interpessoais entre seres humanos, o modelo também pode ser aplicado a relações mais impessoais, como a confiança de um indivíduo em uma organização ou sistema como o de IA. As ambiguidades e contradições inerentes ao âmbito geral da confiança humana em sistemas de IA tornam a aplicação do modelo consideravelmente mais complexa do que quando ele é empregado para examinar relacionamentos entre pessoas. Não obstante, neste artigo, será feita uma tentativa de considerar a decisão de confiar com base no DTM.

Hurley identifica dez elementos essenciais da confiança, que ele divide em duas categorias. A primeira categoria compreende três fatores sobre o indivíduo que confia, relacionados à sua disposição básica para confiar: tolerância ao risco, ajuste psicológico e poder relativo. Esses são fatores característicos de uma pessoa, independentemente de uma situação particular ou da parte que recebe a confiança. Sua disposição para confiar com base nessa categoria poderia ser aplicada tanto a um relacionamento romântico quanto a um relacionamento de negócios.

O grau de tolerância ao risco de uma pessoa influi muito em sua disposição para confiar. Em geral, quando o risco é grande, a confiança é limitada. Contudo, os que praticam o comando de missão estão acostumados a confiar até mesmo em situações de alto risco. Quando os comandantes confiam que seus subordinados exercerão a iniciativa disciplinada com base nas ordens de missão, eles o fazem, em parte, porque entendem como os líderes tomam suas decisões. Os líderes são treinados em certas metodologias, como o processo decisório militar e o processo decisório rápido, ambos os quais ajudam na tomada de decisões e explicam como o líder chegou até elas. A linguagem e os processos em comum ajudam os combatentes a confiarem uns nos outros, porque eles são capazes de imaginar os passos que provavelmente foram tomados para chegar a qualquer decisão. Esse tipo de conhecimento compartilhado pelos membros de um grupo também será necessário no relacionamento entre ser humano e máquina.

Evidentemente, a IA apresenta vários riscos ao longo de um espectro de gravidade, dependendo de sua aplicação. Os possíveis riscos incluem falhas inofensivas, a infiltração do sistema por adversários e ações perigosas com consequências letais. Um risco elevado em particular ou combinação de riscos pode não ser obstáculo para um soldado com alto grau de tolerância. Por outro lado, até um risco insignificante poder ser suficiente para impedir que um soldado com baixa tolerância decida confiar na IA.

O segundo fator individual, o ajuste psicológico, refere-se ao grau de equilíbrio de um indivíduo. As pessoas equilibradas costumam se sentir mais seguras em relação a si mesmas e ao mundo à sua volta. Isso leva tanto a uma maior capacidade quanto à rapidez em confiar. Ainda que a instituição militar inclua indivíduos situados ao longo de todo o espectro de ajuste psicológico, ela promove e fornece as oportunidades educacionais e experiências que

O robô vencedor da equipe Kaist, *DRC-Hubo*, utiliza uma ferramenta para perfurar uma parede na final da competição de robótica da DARPA, em Pomona, Califórnia, 4 Jun 2015. (Foto cedida pela DARPA)

possibilitam um melhor ajuste entre seus membros. O treinamento resulta em maior autoconfiança. A uniformidade ajuda a diminuir as inseguranças raciais e socioeconômicas, questões que podem prejudicar um ajuste positivo fora da organização. A rápida aceitação e adoção de novas missões, equipamentos e integrantes da equipe são valorizadas. Tudo isso contribui para um melhor ajuste psicológico individual, que ajudará na integração da IA.

Embora o ajuste psicológico dos indivíduos de gerações mais jovens varie tanto quanto nas anteriores, está evidente que os futuros soldados em curto prazo se sentem, em geral, mais à vontade com a integração da tecnologia. Isso se deve à difusão da tecnologia, que passou a integrar a estrutura da experiência humana no século XXI. A afinidade da geração Z com a tecnologia está bem documentada[12]. Seus integrantes nasceram em um mundo de farta tecnologia, assimilando-a durante todo o seu desenvolvimento. Como a IA se tornará mais difundida em aplicações civis, os soldados do futuro serão mais propensos a entrar na força com o ajuste psicológico necessário para confiar em tais sistemas. Suas experiências e grau de confiança em relação a aplicações militares de IA dependerão de suas vivências como civis. É concebível que, daqui a uma geração, a questão da confiança do combatente humano na IA já terá sido, essencialmente, resolvida na sociedade.

O último fator individual, o poder relativo, ajuda a determinar a disposição de um indivíduo para confiar com base em seu grau de poder sobre a parte que deve receber a confiança. Os indivíduos que detêm considerável poder com base em sua função dentro de um grupo são mais propensos a depositar confiança nos outros, por estarem aptos a punir os que a violem ou a modificar, e até terminar, o relacionamento com eles. Se os regulamentos e políticas relativos à IA codificassem a supremacia universal dos combatentes humanos sobre tais sistemas, aos militares estaria assegurado um poder relativo que lhes

possibilitaria maior confiança. Caso concedam à IA a capacidade de operar ou agir em qualquer circunstância sobrepondo-se à vontade de um integrante humano da equipe, o poder relativo passará a ser situacional, ficando mais difícil confiar.

Conforme mencionado na introdução, existe um consenso em relação à subordinação da IA aos combatentes humanos, bem como grande cautela quanto à substituição de seres humanos por esses sistemas em decisões com efeitos letais. Essa postura é tranquilizadora, conforme as Forças Armadas avançarem rumo ao futuro. É uma postura que oferece aos militares, individualmente, uma vitória imediata para o fator do poder relativo. Contudo, conforme aumentar a integração da IA, haverá consequências imprevistas, que poderão alterar a dinâmica de poder relativo. Por exemplo, se um ser humano cancelar um esforço de IA e isso resultar em fratricídio ou danos colaterais que não teriam ocorrido sem o cancelamento, haverá uma reavaliação da dinâmica de poder entre humanidade e máquina? Talvez o êxito no emprego de IA em equipes de combate lhe confira uma posição mais elevada de poder relativo, que lhe é negada nos estágios iniciais de integração. Poderá chegar um momento em que o valor da capacidade de IA ultrapasse as preocupações humanitárias dos combatentes humanos, afetando, assim, o fator do poder relativo para a decisão de confiar.

A segunda categoria de Hurley no DTM consiste em sete fatores situacionais que podem ser influenciados pela parte que recebe a confiança de quem a deposita: segurança situacional, semelhanças, interesses, preocupação benevolente, capacidade, previsibilidade/integridade e comunicação. Pode ser útil ter a flexibilidade para avaliar esses fatores identificando a parte que recebe a confiança como a IA por si só ou, às vezes, como uma combinação do sistema de IA, seus criadores e formuladores de política que influenciam sua implementação. Isso se deve ao fato de que a IA Estreita, por carecer de intenção e autoconsciência, pode ser intencionalmente projetada de modo a impedir que ela aja fora dos parâmetros estabelecidos pelos criadores do sistema. Por exemplo, quando se consideram os interesses como um fator situacional na decisão de confiar, eles podem ser, predominantemente, um reflexo do que os criadores do sistema projetaram.

A segurança situacional, capacidade e previsibilidade são todas expectativas comuns no uso de qualquer reforço por máquinas. A segurança situacional é intimamente ligada ao fator de tolerância ao risco na disposição para confiar. Como existe um risco relacionado ao uso de IA em aplicações militares, é importante que ela ofereça segurança situacional, o oposto do risco. Há um certo grau de risco simplesmente porque os pesquisadores e, portanto, os usuários não entendem como a IA processa as informações e chega a uma conclusão. Essa realidade fascinante tem atraído considerável atenção. Em parcerias dentro do ecossistema de ciência e tecnologia, a Agência de Projetos de Pesquisa Avançada do Departamento de Defesa dos EUA está investindo fortemente na IA Explicável. O objetivo dessa tecnologia de IA de "terceira onda" é "criar uma série de técnicas de aprendizado de máquina que produzam modelos explicáveis, ao mesmo tempo que conservam um alto grau de precisão na previsão, de modo que os usuários humanos entendam, confiem adequadamente e administrem efetivamente a geração de parceiros artificialmente inteligentes que está surgindo"[13]. Consiste em uma tentativa de preencher a lacuna entre as decisões ou recomendações feitas por IA e a capacidade de o usuário humano entender por que o sistema chegou a tal conclusão. O êxito no campo da IA Explicável aumentará significativamente a segurança situacional oferecida por esses sistemas aos seres humanos.

Os fatores de capacidade e previsibilidade andam de mãos dadas no campo da tecnologia e são muito simples de entender no relacionamento com a IA. Trata-se de uma questão de competência do sistema. A IA pode cumprir as funções anunciadas? Ela realmente ultrapassa a capacidade humana nas áreas de análise de informações, elaboração de linha de ação ou identificação de alvos? A experiência com IA provavelmente levará os usuários a reconhecerem que ela pode executar as tarefas paras as quais foi projetada com previsibilidade, demonstrada pela raridade de falhas ou desvios de uma norma. A sociedade está, de modo geral, convencida da superioridade das máquinas sobre os seres humanos em várias tarefas. Praticamente ninguém questiona ou verifica à mão os resultados de calculadoras, porque elas foram utilizadas trilhões de vezes para calcular problemas matemáticos sem nenhuma falha. O teste de sistemas antes da implementação pode assegurar a capacidade e a previsibilidade. Se, após acionado, um sistema de IA puder demonstrar sua capacidade para operar da mesma forma, sem erros e segundo as funções definidas,

ele influenciará positivamente a capacidade de confiar do combatente humano.

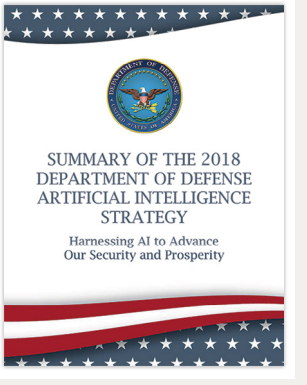
Os fatores restantes — semelhança, interesses, preocupação benevolente e comunicação — são bem mais difíceis de analisar no relacionamento entre um combatente humano e um sistema de IA. A semelhança e os interesses em comum entre homem e máquina são difíceis de estabelecer. Essa pode ser uma área em que tentativas de criar sistemas de IA com interfaces antropomórficas beneficiarão muito a decisão de confiar. O estabelecimento de vínculos com um sistema de IA provavelmente será mais fácil se ele tiver uma aparência ou modo de comunicação semelhante. Um estudo realizado em 2018 sobre interações humanas com um robô demonstrou a capacidade do ser humano para estabelecer vínculos com máquinas com a aparência e comportamento de seres humanos[14]. No estudo, alguns participantes interagiram socialmente com um robô, enquanto outros interagiram com ele de um modo meramente funcional. No final de algumas interações, o robô implorou para não ser desligado. Os participantes que ouviram esse apelo geralmente trataram o robô como se fosse gente. O estudo concluiu que as pessoas são propensas a tratar uma máquina com atributos autônomos mais como um ser humano e menos como uma máquina ou sistema sem tais atributos. Há uma probabilidade maior de que os sistemas de IA desenvolvidos com alguma capacidade antropomórfica promovam a confiança.

É possível que a semelhança e os interesses alinhados possam ser obtidos com o *design* e aplicação da IA Estreita a tarefas de combate, seu objetivo inerente. Se os soldados utilizarem, no nível tático, um sistema de IA que tenha sido criado ou modificado para aplicações táticas, o sistema estará, então, demonstrando semelhança aos combatentes que operam em ambientes táticos. Um futuro sistema de IA Geral poderá experimentar um sentido de consciência de que ele existe e até de que ele deseja lutar e vencer as guerras de nossa nação. Essa seria uma clara demonstração de semelhança e alinhamento de interesses com os combatentes humanos.

Talvez seja possível desenvolver ambientes de treinamento que forjem laços de equipe entre os integrantes humanos e de IA. O Exército está acostumado a receber pessoas heterogêneas e transformá-las todas em militares. A semelhança e o alinhamento de interesses são normalmente obtidos por meio da instrução básica inicial. Os diversos recrutas, oriundos de inúmeras "tribos", criam vínculos por meio das experiências vividas no treinamento, tornando-se parte de uma nova "tribo". Embora a diversidade continue a existir, os soldados apresentam semelhanças significativas e interesses em

O Diretor de Tecnologia da Informação do Departamento de Defesa Dana Deasy (***centro***) e o Major Brigadeiro John N. T. Shanahan, Diretor do Centro Conjunto de Inteligência Artificial (***não incluído***), realizam uma reunião de mesa redonda sobre a estratégia de IA no Pentágono, em Arlington, Estado da Virgínia, 12 Fev 2019. (Foto da Sgt Amber I. Smith, Exército dos EUA)

O documento *Summary of the 2018 Department of the Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity* ("Resumo da Estratégia de Inteligência Artificial de 2018 do Departamento de Defesa: Explorando a IA para Promover Nossa Segurança e Prosperidade"), divulgado pelo Centro Conjunto de Inteligência Artificial, descreve a abordagem e metodologia do Departamento para acelerar a adoção de capacidades de IA, com o intuito de fortalecer nossas Forças Armadas, aumentar a efetividade e eficiência de nossas operações e aumentar a segurança de nossa nação. Para acessar essa publicação, visite https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF.



comum. A confiança é um importante subproduto dessa instrução e experiências de formação. Os seres humanos que treinarem ao lado de sistemas de IA talvez adquiram esse mesmo subproduto.

O fator de preocupação benevolente é a capacidade da IA para colocar as necessidades dos seres humanos acima das suas. É absolutamente necessário que a IA demonstre o entendimento de que os combatentes humanos valem mais do que qualquer componente não humano de uma equipe. A IA se destruirá caso descubra que foi *hackeada* por um adversário? A IA sacrificará sua existência para preservar a vida de seus colegas humanos? Até o ser humano, muitas vezes, decide se importar mais consigo mesmo do que com os que estão à sua volta, e frequentemente aceitamos esse egoísmo em ambientes hipercompetitivos. Contudo, o serviço abnegado é a marca do serviço militar, devendo, portanto, ser exigido da IA. Da mesma forma que os cães de guerra, a IA deve ser capaz de agir corajosamente em defesa de outros combatentes e da missão.

Os futuros sistemas de IA Geral, as máquinas sencientes, provavelmente terão a capacidade para o tipo de coragem que os seres humanos exibem. A coragem física e moral é um valor essencial para os militares e um multiplicador na execução de ações violentas em apoio a objetivos estratégicos, operacionais e táticos. Embora as equipes coesas sejam desenvolvidas com base na confiança mútua resultante, principalmente, de ações cotidianas, os grandes gestos, como os atos de bravura, promovem a confiança e a estima entre seus membros[15]. Durante ações de combate, os militares se inspiram, constantemente, nos atos corajosos de seus companheiros para realizarem mais no campo de batalha do que seria possível de outra forma. A bravura pode se tornar o instrumento para acabar com um impasse, superar uma derrota iminente e sobrepujar uma força inimiga com a violência de ação. A IA Geral pode se portar de um modo que verdadeiramente conquiste a plena confiança de seus companheiros humanos.

Por fim, o elemento da comunicação afeta a maioria dos outros fatores situacionais. Uma boa e frequente comunicação é necessária para desenvolver a confiança. A comunicação com a IA será, sem dúvida, situacional. Conforme explicado anteriormente, é difícil comunicar o processo decisório da IA aos seres humanos, um problema que se busca resolver com a IA Explicável. Os sistemas de IA precisarão de uma interface intuitiva, que promova a comunicação com seus usuários. Se houver algum momento em que a IA passe a impressão de estar evitando a comunicação ou ocultando informações dos combatentes humanos, isso prejudicará a confiança de forma possivelmente irreversível. A comunicação frequente e transparente de sistemas de IA com os soldados ajudará a fomentar o desenvolvimento e manutenção da confiança.

Recomendações

A recém-estabelecida Força-Tarefa de IA do Exército (*AI Task Force — A-AI TF*), sob o Army Futures Command, foi um importante passo com respeito ao desenvolvimento e implementação militar desses sistemas[16]. Não se sabe quais questões éticas, se houver, estão sendo estudadas a fundo como parte dos projetos da A-AI TF. Em cooperação com as atividades da A-AI TF, o Exército pode acelerar a prontidão dos combatentes humanos para confiar na IA de quatro formas. Primeiro, a força precisa conhecer melhor os tipos de sistema em desenvolvimento e suas aplicações previstas nos níveis estratégico, operacional e tático. O sigilo inerente ao desenvolvimento de IA no contexto militar dificulta essa possibilidade, mas deve haver um meio de promover algumas das suas aplicações planejadas. Não basta proclamar: "A IA está chegando". A A-AI TF e outras organizações relacionadas devem buscar formas de comunicar suas atividades ao público mais amplo do Exército dos EUA.

Segundo, A-AI TF deve estudar os fatores de confiança que possibilitam a decisão individual de confiar no que tange aos sistemas de IA. Deve buscar responder, por meio de análises psicológicas, se a atual força tem a disposição necessária para confiar em sistemas de IA como ferramentas ou integrantes de equipes de combate. Os resultados devem ser publicados, acompanhados de recomendações sobre como desenvolver a confiança na IA.

Terceiro, a doutrina do comando de missão deve incluir o conceito de confiança entre seres humanos e sistemas, especialmente os sistemas de inteligência artificial autônomos. Da mesma forma que a doutrina descreve a confiança humana necessária para desenvolver equipes coesas, ela deve detalhar a confiança necessária nos sistemas de IA como parceiros nessas equipes.

Por último, todo soldado deve começar a avaliar sua própria prontidão para confiar nos sistemas de IA que, em breve, mudarão a forma pela qual combatemos nas

guerras da nação. A integração da IA transformará as futuras equipe de combate, de modo semelhante, em alguns aspectos, aos impactos sociais e operacionais decorrentes da integração de mulheres nas qualificações militares das armas combatentes. Os soldados e comandantes tiveram de internalizar os impactos da integração e fazer decisões e ajustes individuais às novas políticas de treinamento e operações das armas combatentes. Para a integração de IA, será preciso conceder aos militares de todos os escalões tempo, espaço e informações adequadas, para que eles possam se perguntar se estão prontos e aptos a confiar que um sistema execute importantes tarefas em sua equipe de combate.

Conclusão

As futuras operações militares serão caracterizadas pela integração generalizada da IA com os combatentes humanos. Algumas pessoas podem argumentar que a integração será gradativa e que a confiança na IA surgirá naturalmente, como consequência da atual e comum afinidade e preferência pela tecnologia já demonstradas pela sociedade. Mesmo que seu argumento se mostre correto, será importante entender a mecânica de tal confiança. É possível, também, que surja uma operação de combate em larga escala que exija o rápido acionamento de sistemas de IA, o que abalará a coesão da equipe de combate humana. Nesse caso, até uma consciência básica da questão de confiança na IA ajudará a força a superar os novos desafios rapidamente. Valendo-se dos atuais conceitos doutrinários sobre a confiança e de um entendimento dos fatores que levam a uma decisão individual de confiar, a força poderá obter uma prontidão básica para confiar. Além disso, com o estudo contínuo por tecnólogos, especialistas em ética, cientistas comportamentais e demais profissionais interessados que servem na comunidade militar, o Exército poderá alcançar um alto nível de prontidão para confiar na IA em equipes de combate coesas. ■

Referências

1. The White House, *National Security Strategy of the United States of America* (Washington, DC: The White House, Dec. 2017), acesso em 5 ago. 2019, https://www.whitehouse.gov/wp-content/uploads/2017/12/NSS-Final-12-18-2017-0905-2.pdf.

2. Department of Defense, *Summary of the 2018 National Defense Strategy of the United States of America* (Washington, DC: U.S. Government Publishing Office [GPO], 2018), acesso em 5 ago. 2019, https://dod.defense.gov/Portals/1/Documents/pubs/2018-National-Defense-Strategy-Summary.pdf.

3. Department of Defense, *Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity* (Washington, DC: U.S. GPO, 2018), acesso em 5 ago. 2019, https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF.

4. Defense Advanced Research Projects Agency, *Department of Defense Fiscal Year (FY) 2020 Budget Estimates* (Washington, DC: Department of Defense, March 2019), acesso em 5 ago. 2019, https://www.darpa.mil/attachments/DARPA_FY20_Presidents_Budget_Request.pdf.

5. Amir Husain, *The Sentient Machine: The Coming Age of Artificial Intelligence* (New York: Scribner, 2017), p. 107.

6. Andrew Ilachinski, *Artificial Intelligence & Autonomy Opportunities and Challenges* (Arlington, VA: Center for Naval Analyses, October 2017), p. 16-17, acesso em 5 ago. 2019, https://www.cna.org/CNA_files/PDF/DIS-2017-U-016388-Final.pdf.

7. Husain, *The Sentient Machine*, p. 9-48.

8. Army Doctrine Reference Publication (ADRP) 6-0, *Mission Command* (Washington, DC: U.S. Government Printing Office, May 2012 [obsoleto]), par. 2-5.

9. ADRP 1, *The Army Profession* (Washington, DC: U.S. GPO, Jun. 2015 [obsoleto]), par. 3-3.

10. Husain, *The Sentient Machine*, p. 89.

11. Robert F. Hurley, *The Decision to Trust: How Leaders Create High-Trust Organizations* (San Francisco: Jossey-Bass, 2012), ProQuest Ebook Central.

12. "How Generation Z Is Shaping Digital Technology", BBC Future, acesso em 5 ago. 2019, https://www.bbc.com/future/sponsored/story/20160309-youth-connection.

13. *A Review and Assessment of the Fiscal Year 2019 Budget Request for Department of Defense Science and Technology Programs Before the Subcommittee on Emerging Threats and Capabilities Armed Services Committee, U.S. House of Representatives*, 115th Cong. (14 Mar. 2018) (declaração de Steven Walker, Diretor, Defense Advanced Research Projects Agency), p. 5-6, acesso em 5 ago. 2019, https://docs.house.gov/meetings/AS/AS26/20180314/107978/HHRG-115-AS26-Wstate-WalkerS-20180314.pdf.

14. Aike C. Horstmann et al., "Do a Robot's Social Skills and Its Objection Discourage Interactants from Switching the Robot Off?", PLOS ONE 13, no. 7 (18 Jul. 2018), acesso em 5 ago. 2019, https://doi.org/10.1371/journal.pone.0201581.

15. ADRP 6-0, *Mission Command*, par. 2-5.

16. Mark T. Esper, Memorandum for Principal Officials of Headquarters, Department of the Army, "Army Directive 2018-18 (Army Artificial Intelligence Task Force in Support of the Department of Defense Joint Artificial Intelligence Center)", 2 Oct. 2018, acesso em 5 ago. 2019, https://armypubs.army.mil/epubs/DR_pubs/DR_a/pdf/web/ARN13011_AD2018_18_Final.pdf.