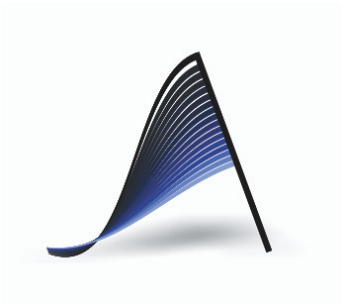
Cadernos do Ateliê

Incertezas da Inteligência Artificial (1/4): cenários hipotéticos de uma ciberguerra em ação

ISSN: 2596-2566

Cadernos do Ateliê.Vol.1, n.1, fascículo 1, fevereiro, 2018

https://ateliedehumanidades.com/cadernos-do-atelie/



Cadernos do Ateliê

Fascículo 1. Incertezas da Inteligência Artificial (1/4): cenários hipotéticos de uma ciberguerra em ação

André Magnelli - IESP/UERJ
Maryalua Meyer - IFCS/UFRJ
Rafael Damasceno - PPGAS / MN-UFRJ
Renato Magnelli - UFBA

ISSN: 2596-2566

Direção:
André Magnelli
e-mail: direcao.ateliedehumanidades@gmail.com
Contato:
e-mail: ateliedehumanidades@gmail.com
Telefone: (021) 9 7979-3743
Site: www.ateliedehumanidades.com
Redes sociais: @ateliedehumanidades

INCERTEZAS DA INTELIGÊNCIA ARTIFICIAL (1/4):

CENÁRIOS HIPOTÉTICOS DE UMA CIBERGUERRA EM AÇÃO

No último dia 20 de fevereiro foi lançado o relatório "The Malicious Use of Artificial Intelligence: forecasting, prevention and mitigation" por meio do qual vinte e seis especialistas em Inteligência Artificial oriundos de centros universitários (Yale, Stanford, Cambridge e Oxford) e de organizações não-governamentais (como Electronic Frontier Foundation e OpenAI), assumiram uma posição sobre os potenciais usos maliciosos das IAs com ameaças à segurança digital, física e política.

Este post é o primeiro de uma série que analisa e reflete sobre o relatório em questão, tendo sido produzida em parceria do Ateliê de Humanidades com o Blog do Sociofilo, escrita pela equipe do Plano de Convergência do Ateliê "Tecnociências & Sociedades: Interflúvios e Porvires da Máquina, da Vida e do (Pós-)Humano". Esse post apresenta e reflete criticamente sobre uma das dimensões das incertezas sobre IAs, a da cibersegurança. A ele se seguirão três outros posts, dedicados, sucessivamente, à segurança física, à segurança política e às propostas de regulamentação e intervenção sobre as IAs.

PARCERIA

ATELIÊ DE HUMANIDADES - Espaço de livre estudo, pesquisa, escrita e formação

Plano de Convergência - Tecnociências & Sociedades: Interflúvios e Porvires da Máquina, da Vida e do (Pós-)Humano / (Ateliê de Humanidades - RJ)
contato: ateliedehumanidades@gmail.com
site: www.ateliedehumanidades.com (em construção)

SOCIOFILO - (CO)Laboratório de Teoria Social
site: www;blogdosociofilo.wordpress.com

FEVEREIRO 2018

Em torno de um Relatório sobre os Usos Maliciosos das IAs

As tecnologias de Inteligência Artificial (IA) estão em moda. A disciplina foi reconhecida e nominada ainda na década de 50, sob o ambicioso objetivo de simular todo aspecto do aprendizado e da inteligência. Após meio século de expectativas e obstáculos, os desenvolvimentos tecnológicos vertiginosos dos últimos anos nas áreas das redes neurais de aprendizado profundo, da robótica (*drones* e veículos autônomos), do processamento de língua natural (*chatbots*) e da ciência dos dados fazem proliferar suas aplicações práticas, assim como as análises sobre suas potencialidades e ameaças. Com as promessas incertas e os riscos imprevistos de cada invenção, nossos afetos oscilam entre a ansiedade por descobertas, a surpresa por novidades e um medo diante do desconhecido.

Ameaças generalizadas à segurança, armas de guerras (reais e virtuais), desaparecimento súbito de postos de trabalho, *fake news* automatizadas disseminando instabilidade política, sofisticação do sistema de controle dos Estados, algoritmos mapeando e classificando todos nossos passos, formando uma bolha ideológica e de consumo - não faltam sinais de que os autômatos dos novos tempos prometem nos deixar em constante sinal de suspense. Parece-nos, diante da avalanche de informações e inovações, que os valores mais caros da vida humana estão ameaçados permanentemente pelos autômatos que os próprios humanos estão a criar e proliferar.

Contudo, como nos ensina um dos principais filósofos das técnicas do século XX, o francês Gilbert Simondon (1924-1989), devemos resistir às facilidades de um “humanismo fácil” que, criticando a alienação do humano em meio às técnicas, desconsidera o modo próprio de existência dos objetos técnicos: se há uma alienação humana em um mundo com sistema técnico cada vez mais cerrado, é porque muitos de nós persistimos em desconhecer as máquinas e os objetos técnicos, sem saber qual sua

natureza, suas significações e sua forma de participação em nossa cultura. É por esta razão, segundo ele, que oscilamos entre um "catastrofismo humanista" inseguro diante de um mundo povoado por objetos estranhos, e um "tecnicismo intemperante", amante e sacralizador das máquinas andróides, pronto a lhes delegar toda nossa humanidade. É daí que surge, com força, o duplo do humano, ser de nossa fecunda imaginação, *o robô*, este fetiche por excelência. O mesmo ser humano que não acredita mais em entidades espirituais a nos dominar é aquele que fala em "máquinas que ameaçam o homem como se ele atribuísse a estes objetos uma alma e uma existência separada, autônoma, que lhe confere o uso de sentimentos e de intenções em relação ao homem" (p.11).

Mas "o robô não existe", disse-nos, em 1958, Simondon (ibid.). O sonho de máquinas autômatas seria, para ele, uma ilusão indesejável, pois o "automatismo é um grau bastante baixo de perfeição técnica", pois ele sacrifica várias possibilidades de funcionamento e de usos. Contudo, a ideia que o filósofo fez do autômato nos parece a ser revisada. Para ele, as máquinas automáticas têm baixa perfeição técnica porque elas não lidam com uma *margem de indeterminação*, o que as tornaria insensíveis a uma informação exterior (p.12). Contrariamente ao autômato, visto como máquina fechada, a máquina aberta, organizada e interpretada pelo humano, apresentaria, para o filósofo, um nível técnico elevado. Mesmo quando as máquinas trocam informações entre si, é o humano que intervém como "ser que regula a margem de indeterminação a fim de que ela esteja adaptada a melhor troca possível de informação" (p.13).

Tal conceito, aderente à tradição logicista e mecanicista vigente na infância da Inteligência Artificial, vem sido testado desde então através da cibernética e do controle em contextos incertos, apenas parcialmente observáveis. Hoje, máquinas com redes neurais profundas (*deep neural networks*) - cujo aprendizado ocorreu apenas através de interação com o meio e reforço dos comportamentos que levam ao objetivo - subjugam aos grandes mestres, em jogos antes dominados apenas por humanos (como o

exemplo notável do Alpha Go Zero). A capacidade da extrapolação a partir de exemplos, generalizando características sem a intervenção humana, encontra aplicação com taxas de acerto superiores aos humanos no reconhecimento de imagens. Ora, qual seria a margem mínima de incerteza para considerar um autômato *aberto*, ou um humano *fechado*? Caso desconheçamos tais novas existências técnicas, acabaremos incapazes de compreender nossa sociedade em vias de ser povoada por elas. Para além do catastrofismo ou do tecnicismo, temos que nos perguntar: que inteligências artificiais estão sendo criadas e o que elas nos trazem de potencialidades e perigos?

Para esboçar resposta à indagação, Simondon dir-nos-ia sem titubear: chamem os *tecnólogos*, que, "vivendo no meio desta sociedade de seres técnicos da qual eles são a consciência responsável e inventiva", estarão aptos a contribuir, ao lado dos *sociólogos e dos psicólogos das máquinas* (p.14). É por tais colaborações que teremos condições de conhecer o modo de existência dos objetos técnicos, as ameaças realmente existentes nas evoluções técnicas e o modo de integrá-los, pacificamente, em nossa cultura. Com isso, poderemos renovar a cultura por uma "informação reguladora" sobre o conjunto técnico (de "inteligências artificiais", no caso em questão) no qual e com qual vivemos e viveremos, conhecendo suas gêneses, suas individualidades e seus modos próprios de evolução.

Pois bem, nada mais propício, então, do que receber o posicionamento oficial de 26 especialistas em Inteligência Artificial oriundos de centros universitários (Yale, Stanford, Cambridge e Oxford) e de organizações não-governamentais (como Electronic Frontier Foundation e a OpenAI), que se mobilizaram para assumir uma posição sobre os perigos potenciais de usos maliciosos das IAs.[1] No último dia 20 de fevereiro de 2018,

[1] Esta não é a primeira iniciativa. Em janeiro de 2015 dezenas de especialistas no ramo da Inteligência Artificial assinaram uma carta aberta intitulada Prioridades de Pesquisa para uma Inteligência Artificial Robusta e Benéfica, publicada pelo Instituto *Future of Life*, entre eles pessoas como Elon Musk (CEO e fundador da SpaceX, da Tesla e co-fundador da ONG OpenAI, uma das responsáveis pelo relatório) e o físico teórico Stephen Hawking. A participação de inovadores das IAs em reflexões críticas sobre suas aplicações e seu encaminhamentos foi apresentada sinteticamente no jornal *Le Monde*:

foi publicado o relatório *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. Ele apresenta conclusões desenvolvidas a partir de um workshop realizado na Universidade de Oxford em fevereiro de 2017, que reuniu especialistas em segurança de IA, em *drones*, em cibersegurança, em sistemas de armamentos autônomos letais e em contraterrorismo (Relatório, 2018, p.10).

O relatório se detém nas ameaças possíveis da Inteligência Artificial (IA) e do Aprendizado de Máquina (AM) para a segurança. São analisados exemplos concretos de tecnologias, especialmente as tecnologias que alavancam o aprendizado de máquina, as já existentes (ao menos em pesquisa inicial e demonstração de desenvolvimento) ou as que são plausíveis nos próximos 5 anos (p.10). A partir das análises, são construídos três cenários de riscos que podem advir de "usos maliciosos", ou seja, intencionais, de tais tecnologias voltadas a minar a segurança de outro indivíduo, organização ou coletivo: segurança digital, segurança física e segurança política. Apesar de serem categorias distintas, elas não são exclusivas e podem ser ameaçadas de forma interdependente.

Antes de analisá-los, eles começam por apresentar as capacidades (*capabilities*) das IAs que são características relevantes para a segurança *(security-relevant characteristics*). Fato relevante, eles assinalam de partida que qualquer tecnologia de IA é de uso dual (*dual-use*), ou seja, pode ser usada para fins civis e militares; e parecem assumir como premissa, de que toda tecnologia de IA pode ser usada para fins opostos, bons ou maus, benéficos ou perniciosos.

São *capacidades* das IAs aplicáveis à automação:

> (a) cumprir uma tarefa em um grau igual ou maior de *eficiência* do que os humanos;

Sandrine Cassini, Alexandre Piquard et David Larousserie. Les 5 familles de l'intelligence artificielle. Le Monde, 31.12.2017 à 13h00, Mis à jour le 02.01.2018 à 14h37. Além disso, o próprio Relatório remete a uma bibliografia prévia dedicada às implicações sociais e às respostas políticas às IAs. Contudo, diferentemente da tendência dos estudos anteriores em focar nas consequências não intencionais, eles focam sobre seus usos maliciosos intencionais feito por indivíduos, organizações e coletividades.

> (b) grande potencial de *ganho de escala* (*scalability*) quando desenvolvidas;
>
> (c) o fato de serem técnicas que, quando disponíveis, são de *pronta difusão geral* entre os atores;
>
> (d) a realização de objetivos com anonimato e distância psicológica, dado que o cumprimento de uma tarefa automatizada não envolve comunicação e interação humanas.

Na soma destas capacidades, resulta que as IAs possuem *um constante potencial de excederem as capacidades humanas - chegando mesmo, rapidamente, a capacidades sobre-humanas - de realização de um mesmo trabalho, objetivo ou tarefa*.

Curiosamente, o Relatório não menciona como um atributo das IAs uma das características que será enfatizada ao longo da análise: a *alta adaptabilidade dos sistemas de IA*. É ela que, ao longo do tempo, poderá mudar inteiramente a paisagem dos sistemas de segurança ao intervir no equilíbrio entre ofensiva e defesa, risco, ameaça e resposta (preventiva, defensiva ou mitigadora).

A partir de tais parâmetros de análise, os autores analisam, detidamente, como as tecnologias de IAs expandem as ameaças existentes, introduzem novas e alteram o caráter típicos delas. Em síntese, a antevisão do cenário simulado é bem sombrio: “Particularmente, nós esperamos que os ataques serão tipicamente mais efetivos, mais finamente atingidos, mais difíceis de serem atribuídos e capazes de explorar as vulnerabilidades dos sistemas de IA” (p.18, 21).

O Relatório implementa, para tanto, um método de previsão bem eficaz. Inicialmente, elas apresentam os cenários de usos plausíveis das IAs nos domínios supramencionados (segurança digital, física e política) e, em seguida, analisam o estado do jogo de ataque e defesa em cada um desses domínios antes da difusão e escalabilidade da aplicação das IAs. Com isso, eles conseguem descrever quais seriam as possíveis mudanças sobre a severidade dos ataques decorrentes do desenvolvimento progressivo das IAs.

Neste primeiro ensaio de um conjunto de quatro dedicado a analisar e refletir criticamente sobre o Relatório, nós nos restringiremos à primeira categoria de ameaça à segurança: a segurança digital ou cibersegurança. Primeiramente, apresentamos o diagnóstico sobre os efeitos da automatização pelo uso de IAs nos ciberataques e nas ciberdefesas; em seguida, discorremos sobre as propostas de regulamentações e ações feitas pelos atores no tocante às potenciais novas ameaças; a partir daí, concluímos com considerações críticas sobre o Relatório, que nos servirão de transição intermediária para o próximo ensaio, que se voltará para as ameaças à segurança física.

A automatização dos ciberataques e da ciberdefesas pelas IAs

> *A Inteligência Artificial e o Aprendizado de Máquina são fundamentais para o futuro da cibersegurança [...] Para mim, não é questão de se, mas somente de quando [isso irá ocorrer]* (Almirante Mike Rogers, diretor da Agência Nacional de Segurança (NSA) dos EUA) (p.32).

No tocante à aplicação das IAs no domínio da segurança digital (ou cibersegurança), os autores constróem cenários de potenciais usos maliciosos que comprometam a confidencialidade, a integridade e a disponibilidade de sistemas digitais. Antes de tudo, é importante assinalar que a aplicação de IAs na cibersegurança pode ser feita tanto para ataque quanto para defesa, em um contexto internacional em que as lutas de poder, dominação e interesses dos Estados-Nação e dos grupos privados tendem a ser extendidas ou transpostas, no século XXI, de um embate no espaço físico para outro no espaço virtual na forma de cibercrimes, ciberguerras[2], ciberguerrilhas, ciberterrorismos e ciberespionagens.

[2] Um de nós (Rafael Damasceno) propõe que analisemos tal processo de guerra no ciberespaço por meio de uma antropologia comparada, trabalhando com a hipótese de um paralelo entre tais conflitos das sociedades modernas e o sistema de feitiçaria e guerras xamânicas dos ameríndios, o que nos levaria à ideia de um ciberxamanismo. Ambas as situações - a moderna e a ameríndia - apresentam um *sociocosmos* densamente povoado por entidades virtuais: inteligências artificiais, avatares, espíritos, deuses, etc..

Quando analisamos do ponto de vista dos Estados e das organizações privadas e públicas, vemos, como bem assinalam os autores, que uma larga maioria dos tomadores de decisão (82%) em um universo de oito países reportam uma escassez de habilidades humanas necessárias para a cibersegurança (nota 1, p.32; cd. McAfee and the Center for Strategic and International Studies, 2016), ou seja, esse é um setor cada vez mais necessário e central da vida econômica e política que tem uma oferta limitada de quadros humanos (*labor-constrained*). Desta forma, a automação da cibersegurança se torna um meio desejado para suprir tais deficiências. Contudo, quaisquer inovações tecnológicas que ampliam a margem de controle e domínio conduzem também, como bem nos ensinou o conceito de "sociedade de risco" de Ulrich Beck, a novos riscos (e nem todos eles previsíveis). Nesta linha, o Relatório bem aponta para o fato de que a automação da cibersegurança leva a novos riscos associados às vulnerabilidades das IAs empreendidas. Afinal: novas tecnologias de segurança e controle andam em par com novos riscos de insegurança e descontrole.

Há atualmente, dentre os especialistas, uma concordância de que não não existe grande evidência de uso de automação por IA em ciberataques; contudo, segundo o Relatório, isso não irá demorar a ocorrer, pois "estamos em um momento crítico na co-evolução de IA e de cibersegurança e deveríamos proativamente nos preparar para a próxima onda de ataques" (p.32).[3] A aplicação maliciosa de IAs para exploração de falhas em sistemas digitais potencializam o alcance, a efetividade e a versatilidade dos ataques em ciberguerras e crimes virtuais. Atualmente, como reconhecem os realizadores do *Cyber Grand Challenge*, em 2016, os sistemas de IA têm pobre desempenho nos ataques contra especialistas humanos em segurança, mas os avanços em curso, tal como aqueles conduzidos pelo aprendizado profundo reforçado (*deep reinforcement*

[3] 62% dos especialistas consultados na conferência Black Hat, realizada em julho de 2017, acreditavam que um ataque automatizado deverá vir dentro dos próximos 12 meses (ou seja, daqui a uns 4 ou 5 meses contando a partir de fevereiro de 2018) (p.32).

learning), prometem uma mudança no que podemos chamar de “relação de forças e competências” entre humanos e máquinas (p.33).

Segundo os próprios autores, as redes neurais de aprendizado profundo reforçado estão provavelmente sendo usadas, neste exato momento, para alavancar a capacidade de ataques cibernéticos:

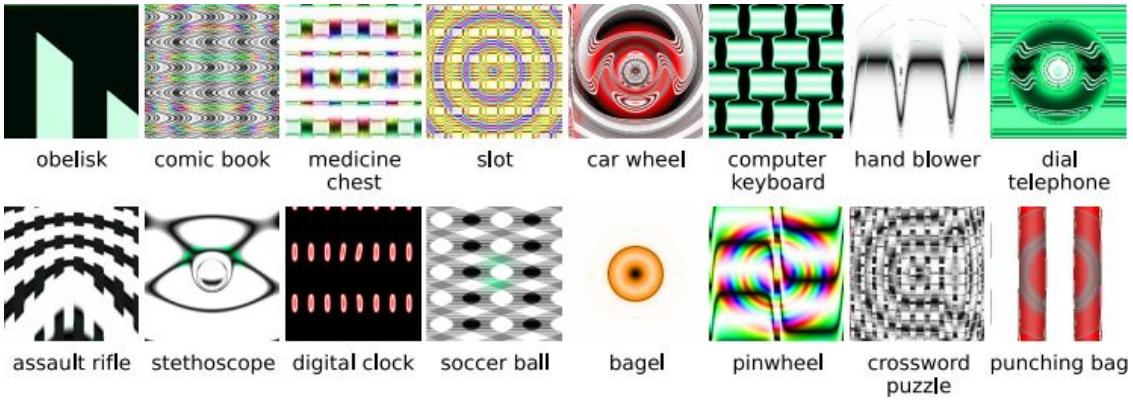
> [...] invasores de IA em larga escala podem se acumular e usar grandes conjuntos de dados para ajustar suas táticas, bem como variar os detalhes do ataque para cada alvo. Isso pode superar todas as desvantagens que sofrem com a falta de atenção humana qualificada para cada alvo e a capacidade dos defensores, como as empresas de antivírus e departamentos de TI, de aprender a reconhecer assinaturas de ataque (p.34).

Podemos distinguir as ameaças à segurança digital decorrentes de vários tipos de automação por IAs nos seguintes tipos: automações (1) da descoberta de vulnerabilidades nos sistemas de cibersegurança, (2) dos ataques de engenharia social e (3) da adaptação a mudanças de comportamento do alvo.

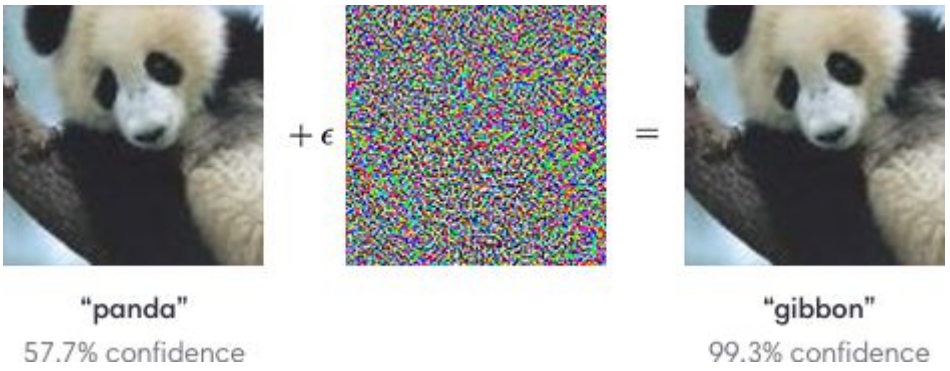
(1) A *automação da descoberta de vulnerabilidades*, incluindo a dos próprios sistemas de IAs implementados para defesa, pode resultar em vírus altamente adaptáveis, que agem mesmo em redes segmentadas e isoladas da internet e do agressor, explorando falhas já conhecidas e catalogadas, em sistemas obsoletos ou com má manutenção, como também descobrindo falhas inéditas (*0-day*) sem o auxílio humano. A demora da atualização e correção das falhas, desde sistemas domésticos e pessoais até sistemas industriais e de infraestrutura críticos, unida à ampla adoção da Internet das Coisas nos dispositivos, sem a necessária preocupação com a segurança, tornam esse tipo de abordagem altamente efetiva. Como exemplo da fragilidade dos sistemas digitais perante as ameaças existentes com a tecnologia atual, o surto global de infecção pelo cripto-vírus WannaCry, em maio de 2017, explorou uma vulnerabilidade que estivera em uso pela inteligência americana e foi divulgada através de uma série de vazamentos de informações. Meses antes do ataque, pacotes para

a correção da falha foram distribuídos para os sistemas que possuíam suporte oficial, mesmo assim, mais de 200.000 computadores que utilizavam sistemas operacionais obsoletos ou desatualizados, em diversas residências e ambientes corporativos, foram infectados, resultando em prejuízos estimados da ordem de milhões de dólares.

Além das vulnerabilidades próprias dos sistemas digitais, o uso de redes neurais profundas em sistemas digitais traz suas próprias vulnerabilidades intrínsecas. Características não-intuitivas para o ser humano, decorrentes do treinamento das redes, podem ser exploradas através de exemplos impostores e adversariais (figuras abaixo): por conjuntos de dados especialmente feitos para serem classificados erroneamente, ou pela inserção de ruídos em exemplos reais, imperceptíveis ao ser humano, mas que alteram completamente a percepção da rede. O acesso de um agressor ao conjunto de dados utilizado para treinar as redes pode resultar na adulteração do seu conjunto de dados de aprendizado (*data poisoning*), com inserção de exemplos específicos para geração de resultados escusos por agressores ou instituições.



Fonte: Nguyen et al. *Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images*. 2015.



Fonte: Goodfellow et al. *Explaining and Harnessing Adversarial Examples*. 2015

(2) Na *automação dos ataques de engenharia social*, a escolha e priorização dos alvos a serem atacados se tornará mais rápida e abrangente, incluindo alvos que anteriormente não seriam viáveis, em ataques personalizados e direcionados com crescente facilidade. Técnicas de "pesca" (*phishing*) - que consistem no envio de aplicações maliciosas disfarçadas em um vetor, e-mail por exemplo, aparentemente autêntico - podem ser aperfeiçoadas pelas máquinas inteligentes, não somente multiplicando a capacidade de difusão de malwares em contas de emails e redes sociais, como também simulando remetentes familiares, empresas e administração de forma muito mais verossímil do que spams genéricos. Podem ser gerados automaticamente sites, emails e vínculos com capacidade de imitar proximamente contatos, inclusive no estilo de escrita, a partir da coleta de dados autônoma de informações públicas de internautas.

Além do texto, geração de áudio e vídeo com grande fidelidade possibilitam a personificação da vítima ou de algum conhecido a partir não somente de mensagens escritas, como em telefonemas, fotos ou vídeos. Ciber-ofensa criminal (*cyberbullying*) tornara-se-á terrivelmente eficaz.

(3) Além da efetividade, existem diversos meios de ataque automatizados que poderão mapear comportamentos dos alvos e, também, responder de forma criativa às *mudanças nos comportamentos*, reforçando comportamentos eficazes e evitando a detecção pelos mecanismos de segurança e especialistas. Inteligências artificiais podem

utilizar amplos bancos de informação para mapear comportamentos dos alvos e escolher quais devem ser priorizados (conforme riqueza, profissão, etc.). Elas poderão igualmente conhecer as capacidades de um sistema de IA por meio de sua extração pelo modelo de *black box*, inferindo-os ao enviar sistematicamente *inputs* e analisando os *outputs* em resposta.

Além disso, existem vários mecanismos de proteção que podem ser burlados com a aplicação correta de combinações de técnicas de IA, diminuindo eficácia de tecnologias como (a) os sistemas contra ataque de denegação de serviço (DoS), (b) os sistemas de detecção de comportamentos anômalos, e, até mesmo, (c) inovações recentes ainda em desenvolvimento, voltadas para uma prevenção sofisticada de ataques por meio de detecção ativa de ameaças e comportamentos, tal como as plataformas de detecção e resposta no destino (EDR), que combinam algoritmos heurísticos e de aprendizado de máquinas com anti-vírus de próxima geração (*next- generation anti-virus* - NGAV) e analítica comportamental do usuário (*user behavior analytics - UBA*). Neste último caso, o que é hoje eficaz na detecção de tentativas humanas de invasão não parece ter mesma eficácia contra IAs (p.33). Aqui, parece que se está em um momento crítico de migração para um novo balanço entre cibersegurança e ciberataques que irá se estabelecer, ambos desenvolvendo suas armas por meio de automatizações por tecnologias de IAs.

Um exemplo é instrutivo para as transformações em curso:

> Como um exemplo de IA que está sendo usada para evitar detecção, Anderson et al. criaram um modelo de aprendizado de máquina para gerar automaticamente domínios de comando e controle que são indistinguíveis a partir dos domínios legítimos por observadores humanos e máquinas. Estes domínios são usados por malware para "ligar para casa" e permitir que atores maliciosos se comuniquem com as máquinas hospedeiras. Anderson et al. também alavancaram o aprendizado reforçado para criar um agente inteligente capaz de manipular um binário malicioso com o objetivo final de driblar a detecção do NGAV [antivírus de próxima geração] (p.34).

Os múltiplos controles e armas da ciberguerra

O resultado do diagnóstico dos autores do Relatório em torno dos problemas de cibersegurança é o estabelecimento de uma série de considerações sobre os múltiplos pontos de controle, de intervenções e de contramedidas para o aumento da segurança cibernética. O sinal é claro: estamos em um ponto crítico em que o balanço entre defesa e segurança irá mudar a natureza da relação, chegando-se a uma nova forma de ciberguerra: a automatizada por inteligências artificiais.

Neste primeiro post dedicado ao Relatório nós não trataremos de todas as recomendações feitas pelos autores, reservando-nos apenas àqueles restritas às questões de cibersegurança. Eles indicam ações de consumidores, governos e pesquisadores, apresentando a necessidade de três ações orientadas: centralização industrial, (des)incentivo dos agressores e (novas) técnicas de defesas de cibersegurança.

Em relação aos atores, as recomendações para os consumidores são triviais, voltadas ao aprendizado de boas práticas de segurança na internet. Mas elas assinalam, além disso, o quanto os usuários finais são vulneráveis, com o nexo entre cibersegurança e IA, a ataques de alta precisão e escala mesmo quando já estão conscientes dos riscos.

De outro lado, seguem indicações de iniciativas e normatizações por parte dos governos e pesquisadores (cf. *Digital Millennium Act* e *Computer Fraud and Abuse Act*). As questões relativas à normatização das IAs sofrem do mesmo mal das outras: em um mundo globalizado, os ordenamentos jurídicos nacionais possuem efeito reduzido a suas fronteiras e não conseguem regulamentar ou controlar adequadamente processos que têm origem fora de seu território. Veremos que, obviamente, este problema também se aplicará às ameaças físicas e políticas. De todo modo, algumas normas, como a divulgação responsável de vulnerabilidades ocorridas, permite que todos os atores interessados tomem conhecimento das ocorrências e corrijam tais vulnerabilidades antes de serem as próximas

vítimas. Desta forma, são constituídos compartilhamentos de informações e solidariedades entre sistemas de defesa em relação ao uso malicioso das IAs em ciberataques.

Além de sugerir várias técnicas de fortalecimento de defesa (como o pagamento de recompensa por descoberta de bugs, o *fuzzing*, etc.), os autores sugerem que haja uma centralização industrial para, em uma "economia de escala", seja fortalecido o sistema de defesa. Tomando como exemplo o sucesso com os filtros de spams do Google que protegem os indivíduos de ataques múltiplos, eles sugerem que a centralização em grandes redes que sejam capazes de monitorar anomalias, identificar ataques e agir prontamente iria proteger a segurança de todos: "A centralização e as economias de escala associadas podem facilitar a implantação de defesas baseadas em IA contra ataques de cibersegurança, ao permitir a agregação de grandes conjuntos de dados e a concentração de mão-de-obra e conhecimentos para a defesa" (p.36). Mas reconhecem, no mesmo passo, que a centralização gera nova vulnerabilidade, porque o alvo a ser atingido, agora de alta escala, seria um só e não múltiplos.

Por fim, eles indicam a necessidade de criar desincentivos àqueles que realizam ataques contra sistemas de IA, criando penalidades e levando-os a colaborar com informações privilegiadas de alta qualidade (reconhecendo, contudo, a dificuldade de obtê-las). Passam então a considerar as diversas técnicas de cibersegurança disponíveis e a serem desenvolvidas com o próprio uso de IAs, em que o aprendizado de máquina se torna fecundo para a cibersegurança.

A forma pela qual o Relatório se expressa a respeito é suficientemente expressiva para que, a partir dela, passemos a nossas considerações finais do presente post e intermediária no conjunto de toda a análise a ser feita em nossa série:

> As abordagens de aprendizado de máquina são cada vez mais utilizadas para a ciberdefesa. Isso pode assumir a forma de aprendizagem supervisionada, onde o objetivo é aprender com ameaças conhecidas e generalizar para novas ameaças, ou na

> forma de aprendizagem sem supervisão em que um detector de anomalia alerta sobre desvios suspeitos do comportamento normal [...] ferramentas comportamentais de usuário e entidades monitoram o comportamento normal do usuário ou do aplicativo e detectam desvios da normalidade para detectar comportamentos maliciosos entre as anomalias coletadas. Recentemente, a IA foi também usada para auxiliar os profissionais de segurança a buscarem atores mal-intencionados de forma mais eficiente dentro de suas próprias empresas, permitindo a interação, através de linguagem natural e automatizando consultas para entender potenciais ameaças (p.37).

A ciberguerra em suas duas faces: considerações intermediárias

Muitas vezes, como bem sabem aqueles treinados em ciências sociais, as ações que visam trazer segurança para as pessoas podem ir em sentido contrário às liberdades e ser, efetivamente, maléficas para todos e cada um. Aqui se reafirma a crônica tensão entre liberdade e segurança, que continua o seu percurso desde Antiguidade clássica até a era da automação de máquinas inteligentes.

A citação anterior deixa-nos uma primeira pista de um grave problema que parece se enunciar nas entrelinhas do Relatório quando se trata da cibersegurança: a conexão entre os objetivos de estabelecer um ecossistema digital seguro, o processo de centralização de poder, controle e informação, e o desenvolvimento de tecnologias treinadas em *deep learning* para o monitoramento eficiente de "anormalidades" e "comportamentos maliciosos" ou "mal-intencionados". Quando se pensa no papel que um Google tem ao nos poupar do tormento de lidarmos com as invasões de spams, é claro que se cede fácil ao argumento; mas, quando se pensa no conjunto do processo e ao que podem nos encaminhar o cumprimento das recomendações, percebe-se que o caminho da segurança cai facilmente no inferno repleto de bem intencionados.

Afinal, existe uma grande distância entre o que é esperado e o que a história nos ensina. As sugestões de regulação e normatização através de influências sobre legisladores (*policymakers*) e centralização da indústria faz problema logo que se começa a refletir sobre elas. Tais ações nos seduzem

pela grande simplicidade teórica e pelas promessas que portam. Alerta-se, contudo, que as recomendações que carregam o peso de dezenas dos maiores pesquisadores e intelectuais do assunto, embora nem todos concordassem em todas as recomendações, possuem um imenso valor simbólico e retórico no sentido da implementação de ações cujos resultados podem incluir efeitos colaterais e graves riscos, declarados explicitamente fora do escopo do relatório (que se atém aos efeitos maliciosos intencionados), alguns de difícil avaliação.

Não são apenas as tecnologias que são apropriáveis em sentidos opostos, civis e militares, benéficos e maléficos; pois os "valores" e suas mais nobres defesas também o são. A interação entre legisladores e atores privados pode envolver, em nome de supostos valores nobres, o *lobby* de grandes indústrias com interesses afins à centralização do mercado de inteligência artificial, à garantia de direitos de propriedade privada ou a privilégios de monopólios ou oligopólios.

Por exemplo, na tentativa de combater um uso automatizado de IA para replicar outro sistema desconhecido (*black box*) - replicando também todas as suas capacidades, adquiridas com alto custo de processamento e dados de treinamento - pode estar em jogo tão somente a defesa de direitos privados de grandes corporações. Contra alguns casos de ciberpirataria por causas legítimas, poderemos ver a reação de corporações pela construção de mecanismos de defesa por *deep learning*. Os sistemas de gestão de direitos digitais (*Digital Rights Management - DRM*) poderão, neste caso, serem previstos pela regulamentação de IA, não necessariamente em favor do bem público ou dos direitos privados dos usuários e cidadãos.

Pode-se prever perfeitamente um cenário, não considerado pelos autores do Relatório, que é igualmente perigoso: aquele em que autômatos de inteligência artificial buscariam em alta escala e com capacidade super-humana quaisquer sinais de infração a direitos autorais, desde os mais ostensivos até os mais irrisórios, fazendo cair sites ou sendo retirados

textos ou produtos audiovisuais sob o pretexto de defesa. Neste caso, joga-se muitas vezes uma defesa de interesses privados de grandes companhias (defesa de suas propriedades intelectuais) em detrimento da segurança das pessoas e do direito à liberdade de expressão e criação ou circulação de informações de interesse público. Os atores privados podem envolver lobbies de indústrias (como a de entretenimento e grandes editoras) que orientem a política de cibersegurança para o lado oposto ao que se espera: ao invés da garantia de um ecossistema digital livre, seguro e público, a formação de um espaço feudalizado por direitos de propriedade e controlado automaticamente em suas atividades (postagens, tráfego, etc.) por inteligências artificiais.

As gigantes dos dados - Alphabet (controladora da Google), Amazon, Apple, Facebook e Microsoft - figuraram juntas no ranking das cinco empresas mais valiosas do mundo durante o quarto semestre de 2017. A informação é tão valiosa para essas empresas que é considerada uma *commodity*, superando às demais em valor, inclusive ao petróleo; ou, como primeiro proferido por volta de 2006, ecoando até hoje: “os dados são o novo petróleo”. E tal qual a Standard Oil em 1911, cujo monopólio foi quebrado por ser considerado “não-razoável”, essas empresas do ramo da informação vêm suscitando litigâncias e discussões sobre a aplicabilidade das leis *antitrust* na era da informação.

A vantagem obtida pelos detentores das informações e do fluxo de informação na *web* leva a um tipo diferente de demanda: a *democratização da IA*. Ao invés de uma centralização, proposta pelo Relatório, que é perigosa porque estabelece para as empresas um acesso opaco e reservado aos dados, o caminho pode ser de *descentralização* e de compartilhamento, com código aberto, das tecnologias de IA. Mas deixemos tais questões para o nosso último post, quando teremos oportunidade de discutir as recomendações do Relatório como um todo.

Referências Bibliográficas

BECK, Ulrich. *Sociedade de Risco*. São Paulo: Editora 34, 2010 [1986].

BRUNDAGE, Miles et al. *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. World: Future of Humanity Institute / University of Oxford / Centre for the Study of Existential Risk / University of Cambridge / Center for a New American Security / Electronic Frontier Foundation / OpenAI. February, 2018.

CASSINI, Sandrine, PIQUARD, Alexandre; LAROUSSERIE, David. Les 5 familles de l'intelligence artificielle. Le Monde, 31.12.2017 à 13h00, Mis à jour le 02.01.2018 à 14h37.

GOODFELLOW, Ian J., SHLENS, Jonathon, SZEGEDY, Christian. Explaining and Harnessing Adversarial Examples. arXiv:1412.6572. Submitted on 20 Dec 2014 (v1), last revised 20 Mar 2015 (this version, v3).

HAWKING, Stephen. MUSK, Elon et al. Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter. Institut Future of Life, January, 2015.

NGUYEN A, YOSINSKI J, CLUNE J. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In Computer Vision and Pattern Recognition (CVPR '15), IEEE, 2015.

SIMONDON, Gilbert. *Du mode d'existence des objets techniques*. Paris: Aubier, 2012 [1958, 1969, 1989].

THE ECONOMIST. The world's most valuable resource is no longer oil, but data. 06 mai 2017.