# Conselhos anônimos: Se você quiser reduzir os riscos da IA, você deve assumir funções que melhoram as capacidades da IA?

Por Benjamin Hilton · Publicado em outubro de 2022

Já argumentamos que a prevenção de uma catástrofe relacionada à IA pode ser o problema mais premente do mundo e que se por um lado o progresso da IA nas próximas décadas poderia resultar em benefícios enormes, por outro ele poderia impor riscos graves, talvez até existenciais. Por isso, achamos que trabalharmos com algum aspecto da pesquisa técnica sobre a IA — pesquisa relacionada à segurança da IA — poderia ser um plano de carreira de impacto particularmente alto.

Mas há muitas maneiras de conduzir esse plano que envolvem a pesquisa sobre as *capacidades* da IA ou o melhoria das capacidades em áreas além da pesquisa – isto é, tornar os sistemas de IA melhores em habilidades específicas – em vez de limitar seu trabalho ao território da *segurança*. Resumidamente, isto é porque:

- O trabalho com as capacidades e algumas formas de trabalho com a segurança estão entrelaçados.
- Muitas maneiras de entendermos o suficiente sobre a IA para contribuirmos à segurança são funções de melhoria de capacidades.

**Então, se você quiser ajudar a prevenir uma catástrofe relacionada à IA, você deve considerar a possibilidade de exercer funções que também melhoram as capacidades da IA ou deve evitá-las?**

Achamos que esta é uma questão difícil! As funções que envolvem a melhoria das capacidades poderiam ser benéficas ou prejudiciais. Com qualquer função, existe uma variedade de considerações – e pessoas sensatas discordam se, e em que casos, os riscos são maiores do que os benefícios.

Então, pedimos que 22 pessoas que consideramos as mais bem informadas sobre esta questão – e que, sabíamos, tinham opiniões divergentes — escrevessem um resumo de seus pensamentos sobre o assunto. Recebemos 11 respostas muito interessantes e achamos que provavelmente representam razoavelmente bem a variedade de pontos de vista do conjunto mais abrangente de pessoas.

Esperamos que estas respostas ajudem a informar pessoas que estão tomando decisões difíceis sobre exercerem funções que possam melhorar as capacidades da IA.

Se não for possível você seguir algumas das recomendações abaixo, não se preocupe! Confira o nosso perfil de problema sobre como prevenir uma catástrofe relacionada à IA, onde são apresentados os termos, conceitos e argumentos mencionados aqui.

As sugestões abaixo foram escritas por pessoas cujo trabalho respeitamos e que nos pediram que não revelássemos sua identidade. **Estas citações não representam as opiniões de 80,000 Horas e em alguns casos as sugestões oferecidas podem contradizer as nossas explicitamente.** Mesmo assim, achamos que é importante apresentarmos a variedade de opiniões sobre assuntos difíceis quando houver desacordo entre pessoas sensatas.

Incluímos abaixo as respostas desses 11 especialistas em sua totalidade. Fizemos apenas algumas pequenas modificações para melhorar a clareza e facilidade de leitura.

## Os conselhos de 11 especialistas anônimos

**Especialista no. 1: No momento, eu gostaria que os laboratórios de IA fossem mais devagar  na margem, but…**

No momento, eu gostaria que os laboratórios de IA fossem mais devagar na margem, mas acho que não é óbvio que o trabalho com as capacidades é líquido negativo e pessoas sensatas podem

discordar sobre este ponto. O que digo abaixo provavelmente seria diferente se eu tivesse uma alta confiança em que é extremamente ruim melhorar as capacidades (principalmente se eu acreditasse que o avanço das capacidades em um mês causa muito mais mal do que o bem que a avanço da pesquisa de alinhamento causa em um mês. Com isso em mente:

- **Os laboratórios de IA são lugares ótimos para desenvolver boas habilidades**, principalmente para pessoas em funções técnicas (engenheiro ou pesquisador de aprendizado de máquinas) em tais laboratórios. Se você estiver em início de carreira e conseguir entrar num desses laboratórios e acha que seria mais interessante e melhor para você a nível de aptidão pessoal do que outras funções que você está pensando em exercer, então você provavelmente deveria trabalhar nessa área – segundo uma gama de opiniões bastante ampla (que inclui a opinião de que o aprimoramento de capacidades é líquido negativo), o investimento no seu capital humano provavelmente gera mais bem do que a sua contribuição na área de capacidades (num escalão mais baixo) causa danos.
- **Vale a pena inquirir especificamente sobre o trabalho em equipes de segurança em empresas de capacidade de IA** e não presumir que você precisa escolher ou uma função relacionada puramente às capacidades ou nenhuma função; isto nem sempre vai dar certo, mas minha expectativa seria de que algumas vezes você vai poder voltar o seu trabalho a projetos de segurança (e isto provavelmente – embora nem sempre – vai tornar seu aprendizado no emprego um pouco mais relevante/útil a projetos de segurança em que você possa trabalhar mais tarde).
- **Existem algumas funções (normalmente de escalão mais alto) que parecem ser uma super alavanca para aumentar as capacidades em geral** e não lhe ajudam a desenvolver as habilidades que são altamente transferíveis a projetos

exclusivamente de segurança – por exemplo, o angariamento de fundos para um laboratório de IA ou o trabalho na área de comunicação de um laboratório de IA que consiste em criar "hype" em torno de seus resultados na área de capacidades. Essas funções serão provavelmente uma má escolha, a menos que você apóie mesmo o laboratório e ache que ele se alinhe bem com suas opiniões sobre a segurança e os seus valores.

- **Presuma sempre que você é afetado psicologicamente pelo ambiente em que trabalha.** Pelo que já observei, as pessoas que trabalham em laboratórios de capacidades tendem a ter or desenvolver sistematicamente opiniões de que o alinhamento de IA será bastante fácil e imagino que isso é em boa parte devido a efeitos de raciocínio motivado e conformismo social. Penso que o que mais me animaria seria um pesquisador de alinhamento de IA que passa um pouco de seu tempo em laboratórios de IA e um pouco fora desses ambientes (ou numa empresa como ARC or Redwood que trabalha com a segurança exclusivamente, um grupo acadêmico focado na segurança, ou fazendo pesquisa independente). Parece que você ganha uma perspectiva importante nesses dois ambientes e vale a pena combater a inércia de permanecer numa função que é confortável, ser promovido continuamente e não chegar nunca a trabalhar naquilo que é mais central à segurança.

**Especialista 2: Existem partes da segurança da Inteligência Artificial Geral (IAG) que estão intimamente ligadas às capacidades…**

Existem partes da Inteligência Artificial Geral (IAG) que estão intimamente ligadas às capacidades, e em particular a amplificação e as variantes, e acho que vale a pena entrar nessas áreas on net (embora sem alto nível de confiança). Isto é, uma mensagem frequentemente expressa hoje em dia é que o pessoal do Altruísmo Eficaz deve trabalhar apenas nas áreas mais puras da segurança, mas acho que isso é menos do que ideal, dado que uma solução completa envolve áreas que incluem as capacidades também.

**Especialista no. 3: Se a humanidade obtiver a inteligência artificial geral bem antes de saber como direcioná-la…**

Se a humanidade obtiver a inteligência artificial geral bem antes de saber como direcioná-la, então é provável que a humanidade causará seu próprio extermínio com ela porque a humanidade não é capaz de se coordenar bem o suficiente para evitar que as pessoas mais erroneamente otimistas lancem uma superinteligência não-amigável antes de qualquer pessoa saber como construir uma amigável. Sendo assim, no ambiente atual – onde as capacidades estão avançando mais rapidamente do que o alinhamento – o efeito de primeira ordem de se trabalhar com as capacidades da IA é acelerar a destruição de tudo (ou, bem, o cone de luz emanando da Terra num futuro próximo). Este efeito de primeira ordem, me parece, domina completamente os vários efeitos positivos de segunda ordem (tais como obtermos maiores conhecimento sobre o estado atual da pesquisa sobre a capacidade e sermos capazes de influenciar culturalmente os pesquisadores da área de capacidades da IA). (Existem também efeitos de segunda ordem negativos em se trabalhar com as capacidades apesar dos efeitos de primeira ordem negativos, como por exemplo como isso enfraquece o efeito cultural positivo que poderia surgir se todas as pessoas conscienciosas se recusassem a trabalhar com as capacidades em vez de com uma melhor história do alinhamento.)

Por outro lado, o caso não é tão simples. Às vezes, a pesquisa voltada ao alinhamento naturalmente impulsiona as capacidades. E muitas pessoas não podem ser persuadidas a abandonar a pesquisa sobre as capacidades, seja qual for o estado da pesquisa sobre of alinhamento. Sendo assim, acrescento que a pesquisa sobre as capacidades pró-sociais é possível, contanto que seja feita estritamente de forma privada, numa equipe de pesquisadores que compreendem o tipo de perigo com que estão se envolvendo, e que seja bem capaz de refrear-se e não implementar sistemas que não

sejam seguros. (Note que se existem múltiplas equipes que creem contar com essa propriedade, o cone de luz à frente é destruído pelas pessoas mais erroneamente otimistas entre eles. A equipe precisa não só tentar alinhar suas IAs mas também ser capaz de detectar quando o sistema implementado não seria amigável; o segundo é uma tarefa muito mais difícil.) Por padrão, as equipes que dizem estar agindo de forma privada, quando isso conta, irão publicar com prazer uma série de pistas que vão levar qualquer pessoa atenta direto às suas conclusões sobre as capacidades (justificadas, talvez, por argumentos como “se não publicarmos, não vamos conseguir contratar os melhores candidatos”), e vão voltar a agir de forma privada somente quando já for tarde demais. Mas se você encontrar uma equipe que se dedica intensamente ao alinhamento, e que já está se recusando a publicar, isso é um pouco melhor, na minha opinião.

Imagino ainda que as pessoas que dizem “não podemos realizar um trabalho efetivo com o alinhamento até termos mais capacidades” estão, embora não estejam completamente erradas, acabando rapidamente com um recurso escasso e necessário. Ou seja: sim, existe um trabalho com o alinhamento que vai se tornar muito mais fácil assim que tivermos verdadeiras IAGs em nossas mãos. Mas também existem obstáculos previsíveis que permanecerão mesmo então e que requerem um período extenso para resolvermos. Se a humanidade conseguir sobreviver 5 anos depois de inventar a IAG e antes que alguém destrua o universo, e existir um problema que nos levará 20 anos para resolver sem um IGA na sua frente e 10 anos para resolver com uma IAG na sua frente, então é melhor começarmos isso agora, e acelerar as capacidades não está ajudando. Então, mesmo se a equipe conduzir seu trabalho de forma realmente privada, meu palpite é que os avanços das capacidades estão consumindo um tempo de que necessitamos. A única área em que é claramente uma boa opção, na minha opinião, melhorar as capacidades é quando essas melhorias de capacidade resultam

necessariamente do avanço em nosso conhecimento sobre o alinhamento da IA em casos em que há o maior número de gargalos em série e onde esses avanços são feitos apenas de forma privada. Mas poucas pessoas podem dizer onde se encontram os gargalos em série e então acho que uma boa regra geral é: não melhore as capacidades; e se você tiver que fazer isso, assegure-se de que está sendo feito de forma privada.

**Especialista no. 4: Há muitas considerações, nos dois sentidos, de dimensões comparáveis…**

Há muitas considerações, nos dois sentidos, de dimensões comparáveis (na minha opinião), então penso que você não deveria confiar inteiramente em nenhuma resposta. Para citar algumas: (1) termos mais pessoas alinhadas adquirindo habilidades relevantes e que poderão mais tarde trabalhar diretamente com a redução dos riscos-x aumenta a qualidade esperada do trabalho com os riscos-x (bom), (2) escalas de tempo mais curtas significam menos tempo para o trabalho com o alinhamento e a governança (ruim), (3) escalas de tempo mais curtas significam menos atores criando IAG em momentos críticos (bom), (4) mais pessoas alinhadas em organizações relevantes podem ajudar a fortalecer a vontade política nessas organizações para abordar questões de segurança (bom), (5) escalas de tempo mais curtas significam menos tempo para pessoas alinhadas ascenderem a cargos de influência em organizações relevantes (ruim), (6) escalas de tempo mais curtas significam menos tempo para a geopolítica mudar (bom ou ruim, não está claro).

Meu principal conselho é não subestimar a possibilidade de mudanças de valores. Não me surpreenderia ouvir uma história de alguém que entrou na OpenAI ou DeepMind para desenvolver habilidades em aprendizado de máquinas, formou amizade com um grupo de pesquisadores de IA, desenvolveu uma boa compreensão dos sistemas de IA que podemos construir e no final acabou declarando estar de acordo com este ou aquele motivo para pensar

que os riscos da IA são exagerados, sem nunca encontrar um argumento a favor dessa conclusão que defenderia a partir de seu posicionamento inicial. Se você for trabalhar numa função relacionada à melhoria das capacidades, quero que você se assegure de que seu círculo de amizades continue incluindo pessoas que se preocupam com os riscos-x da IA e que continuem lendo os trabalhos sendo feitos na área dos riscos-x da IA.

Se eu tivesse que recomendar uma decisão sem ter nenhuma outra informação sobre você, imagino que eu seria (a) a favor de uma função na área de melhoria de capacidades para o desenvolvimento de suas habilidades se você tomou as precauções acima (e contra se você não tomar), (b) a favor de uma função na área de melhoria das capacidades na qual você exerça sua influência para trabalhar com os riscos-x da IA se você estiver nos escalões mais altos (e contra se você estiver num escalão mais baixo).

**Especialista no. 5: Não há, no momento, nenhum plano de impedir que a IAG destrua o mundo…**

Não há, no momento, nenhum plano de impedir que a IAG destrua o mundo. Isso tem sido comparado, de forma justificada e válida, ao filme “Não Olhe Para Cima,” mas existem empresas atraindo o asteróide com a esperança de que vão ter lucros, só que eles não têm nem um plano para evitar que todo mundo morra. Nessas circunstâncias, acho que é inaceitável – em termos de consequências, de deontologia, ou como ser humano – “tocar fogo na praça pública das capacidades” através da publicação dos avanços das capacidades, do acesso a modelos, ao código-fonte, através da atenção chamada às técnicas que você usou de forma fechada para realizar avanços nas capacidades, da exibição de capacidades empolgantes que fazem as pessoas ficarem animadas a entrar no campo ou através de ações que visivelmente façam parecer que as empresas de IA vão ser super lucrativas e patrocináveis e tudo mundo deveria criar uma também.

Existem vagas de empregos em todo lugar na IA, infelizmente. Arranje um emprego numa empresa que não vai romper os limites das capacidades, não vai disponibilizar seu código-fonte, não vai incentivar a entrada de mais profissionais no campo do aprendizado de máquinas; de preferência, obtenha deles uma confirmação clara de que sua pesquisa será sempre fechada e assegure-se de que você não vai trabalhar com outros pesquisadores que vão ficar tristes em não poder publicar trabalhos super interessantes que lhes trariam maior prestígio.

Fechado é cooperação. Aberto é deserção. Assegure-se de que seu trabalho não está contribuindo para a destruição da humanidade pela humanidade, ou não trabalhe.

Também tente não cair nas desculpas esfarrapadas dos sonhos remotos da relevância do alinhamento.

**Especialista no. 6: Acho que a perspectiva simples de que "o trabalho com as capacidades faz a IAG ficar mais próxima, o que é ruim por causa dos riscos-x da IA," provavelmente…**

Acho que a perspectiva simples de que "o trabalho com as capacidades faz a IAG ficar mais próxima, o que é ruim por causa dos riscos-x da IA," provavelmente está direcionalmente correto em média, mas é uma simplificação tão vastamente exagerada que é praticamente inútil como heurística.

Existem muitas maneiras diferentes em que o trabalho com as capacidades pode ter tanto efeitos positivos quanto negativos e estes podem variar muito dependendo tanto da natureza do trabalho quanto da forma como é usado e divulgado. Estas são algumas questões que eu gostaria de considerar ao julgar o efeito concreto do trabalho com as capacidades:

- **Qual é o efeito direto nas escalas de tempo da IAG?** É provável que o trabalho com as capacidades que vai

gradualmente diminuindo os gargalos da IAG (a que me referirei como o "trabalho com os gargalos da IAG") fará com que a IAG chegue mais cedo. A maior categoria que vejo aqui é o trabalho que melhora a eficiência do treinamento dos grandes modelos que têm uma compreensão do mundo, seja através de melhorias da arquitetura, de sistemas de otimização, melhoria do treinamento de precisão reduzida, do hardware, etc. Parte do trabalho deste tipo pode ter um impacto contrafactual menor: por exemplo, pode ser difícil usar esse trabalho como base para outros avanços porque ele é peculiar ao hardware, software ou modelos de hoje, ou ele pode ser semelhante ao trabalho já sendo realizado por outros.

- **Qual é o efeito na aceleração?** O trabalho com as capacidades pode ter um efeito indireto nas escalas de tempo da IAG ao incentivar outros a (a) investir mais em trabalho com os gargalos das capacidades da IAG ou a (b) gastar mais com o treinamento dos modelos grandes, levando a uma escala de tempo de gastos acelerada que no final vai resultar na IAG. Ao mesmo tempo, parte do trabalho com as capacidades pode incentivar outros a trabalharem com o alinhamento, talvez dependendo de como é apresentado.
- **Qual é o efeito na velocidade de lançamento?** Maiores gastos no treinamento dos modelos grandes agora podem levar a um ritmo mais lento do aumento dos gastos na época da IAG através da redução das "verbas excedentes," ou seja, dinheiro que se encontra inerte porque a empresa não sabe onde usá-lo. Isto pode trazer melhores resultados porque daria ao mundo mais tempo com os modelos de quase-IAG, pois tendo esses modelos poderia chamar mais atenção para o alinhamento da IA, torná-lo mais tratável empiricamente, e fazer com que seja mais fácil as instituições se adaptarem. Claro, um gasto maior com o treinamento dos modelos grandes provavelmente envolve um certo trabalho com os gargalos das capacidades da IAG, e os benefícios são limitados pelo fato de que nem todas

as pesquisas de alinhamento requerem os modelos mais capazes e que a comunidade toda de alinhamento da IA está crescendo pelo menos em parte independentemente dos avanços das capacidades.

- **Qual é o efeito nos riscos de desalinhamento?** Parte do trabalho com as capacidades pode tornar os modelos mais úteis sem aumentar os riscos de desalinhamento. De fato, o alinhamento dos modelos grandes de linguagem os torna mais úteis (e portanto pode ser considerado trabalho com as capacidades), mas não dá ao modelo-base uma compreensão não-trivialmente melhor do mundo, o que normalmente é visto como um impulsionador essencial dos riscos de desalinhamento. Esse tipo de trabalho deve reduzir indiretamente o risco de desalinhamento através da melhoria de nossa habilidade de realizar coisas (inclusive de vencer a concorrência com a IA desalinhada, conduzir mais pesquisa sobre o alinhamento e implementar outras mitigações) antes e durante o período de maior risco. Também vale a pena considerar os efeitos em outros riscos tais como o risco do uso impróprio, embora normalmente estes sejam considerados menos existencialmente graves.
- **Qual é o efeito na pesquisa sobre o alinhamento?** Parte do trabalho com capacidades permitiria a realização de um novo trabalho com o alinhamento, inclusive trabalho com esquemas de alinhamento externo que envolvem avaliações assistidas pela IA tais como o debate e o estudo empírico do desalinhamento interno (embora exista um debate acirrado sobre quanto tempo antes da IAG espera-se que o segundo seja possível). Outros tipos de trabalho com as capacidades podem permitir que os modelos assistam ou conduzam pesquisa sobre o alinhamento. Na verdade, grande parte do trabalho com os gargalos da IAG pode se encaixar nesta categoria. Claro, muita pesquisa sobre o alinhamento na atualidade não enfrenta gargalos relacionados com as

capacidades dos modelos, como por exemplo o trabalho teórico e a interpretabilidade.
- **Como o trabalho será utilizado e divulgado?** As desvantagens potenciais do trabalho com as capacidades muitas vezes podem ser mitigadas, talvez completamente, através do uso ou divulgação do trabalho de uma certa maneira ou da sua não divulgação. No entanto, tais mitigações podem ser frágeis e podem também reduzir as vantagens do alinhamento.

No geral, não creio que um projeto poder ou não ser rotulado como "de capacidades" à primeira vista lhe diz muito sobre ele ser bom ou ruim. Creio, sim, que um trabalho com os gargalos da IGA que é divulgado publicamente é provavelmente prejudicial no fim das contas, mas não de maneira óbvia. Já que essa perspectiva depende tanto de decisões difíceis, como por exemplo determinar o valor relativo do trabalho com alinhamento a nível teórico versus empírico, devo ser cauteloso em meu conselho geral com relação a tal trabalho:

- Evite o trabalho com os gargalos da IAG se ele não apresentar uma vantagem clara relativa ao alinhamento ou à mitigação meticulosa, mesmo se houver uma vantagem relativa ao seu aprendizado ou carreira. Observe que eu não consideraria a maior parte do trabalho acadêmico na área de aprendizado de máquinas como trabalho com os gargalos da IAG, já que este não foca em coisas como a melhoria da eficiência do treinamento dos modelos grandes que têm compreensão do mundo.
- No caso do trabalho relacionado à IAG que visa o alinhamento mas também tem impacto nos gargalos da IAG, vale a pena discutir o projeto com outras pessoas antes para verificar se o trabalho vale a pena de forma geral. Minha expectativa é de que o resultado correto da maior parte dessas discussões seria a

conclusão de prosseguir com o projeto, simplesmente porque o efeito de um único projeto que não otimiza uma coisa provavelmente é muito pequeno comparado a um grande número de projetos que estão otimizando essa coisa. Mas o que está em jogo é importante a ponto de valer a pena examinar essas considerações a nível do objeto.

- O trabalho que está relacionado apenas tangencialmente à IAG, como um projeto sobre a teoria do aprendizado de máquinas ou que aplica o aprendizado de máquinas a algum problema da vida prática, merece menos escrutínio do ponto de vista da IAG, mesmo que possam ser chamadas de "capacidades." O efeito de tal projeto é provavelmente dominado por seu impacto no problema de ordem prática e no seu aprendizado, sua carreira, etc.
- Estudantes: não fiquem preocupados. A vasta maioria dos projetos de estudantes acabam não fazendo muita diferença, então você provavelmente poderia escolher um projeto com o qual você mais aprenderia (embora, é claro, seja mais provável que você vá aprender mais sobre o alinhamento se o projeto tiver a ver com o alinhamento.)

**Especialista no. 7: As escalas de tempo são curtas, estamos numa corrida em direção ao precipício…**

As escalas de tempo são curtas, estamos numa corrida em direção ao precipício, e parte das pesquisas sobre a melhoria das capacidades vale a pena se proporcionar uma grande recompensa de alguma outra forma, mas automaticamente você deve ser manter cético.

**Especialista no. 8: No geral, acho que há muitas vantagens em pessoas preocupadas com os riscos extremos/existenciais da IA…**

No geral, acho que há muitas vantagens em pessoas preocupadas com os riscos extremos/existenciais da IA trabalharem em áreas que aparentemente têm como foco principal melhorarem as capacidades,

se existirem outras boas razões pelas quais elas vão tomar essa direção. Isto é porque (1) a distinção entre capacidades e segurança é precária a ponto de muitas vezes não ser útil; (2) minha expectativa é que conclusões relevantes à segurança irão surgir de áreas que hoje podem ser codificadas como capacidades, e então recomendo um portfólio muito mais diverso para casos em que indivíduos preocupados com os riscos da IA possam ampliar suas habilidades; (3) existem benefícios significativos em contar com indivíduos conscientes dos riscos da IA que se destacam em organizações de IA e nos campos do aprendizado de máquinas; (4) o trabalho com as capacidades é e será altamente incentivado, muito além do aumento marginal de 80 mil indivíduos, na minha opinião.

1. A distinção entre capacidades e segurança faz sentido no abstrato e vale a pena prestarmos atenção a ela. Os laboratórios que tentam trabalhar diferencialmente com a segurança e publicam nesta área em vez de trabalhar com as capacidades merecem nosso aplauso. Os patrocinadores filantrópicos e outros atores deveriam ser conscienciosos sobre as maneiras como seus investimentos podem expandir diferencialmente a segurança/alinhamento, relativos às capacidades, ou não. Por outro lado, a meu ver, quando alguém conduz uma avaliação sofisticada, na prática é muito difícil estabelecer uma distinção clara entre a segurança e as capacidades, e então muitas vezes essa distinção não deveria ser usada para orientar a ação. A interpretabilidade, a robustez, o alinhamento de modelos a curto-prazo, generalizações fora da amostra são todas áreas importantes que melhoram de forma plausível a segurança bem como as capacidades. Existem muitas circunstâncias em que até um ganho puro em segurança pode gerar perigo, como por exemplo se este ocultar dados indicativos de riscos posteriores ao alinhamento ou incentivar atores a implementar modelos que em outras circunstâncias seria arriscado demais implementar.

2. A meu ver, o campo da segurança da IA/IAG passou por um processo de expansão que resultou num número maior de abordagens que anteriormente eram consideradas distantes demais dos riscos mais extremos. A interpretabilidade mecanista ou o alinhamento dos modelos de aprendizado profundo existentes são considerados por muitos hoje uma aposta valiosa na segurança a longo prazo, enquanto que vários anos atrás estes estavam muito mais na periferia das considerações. Minha expectativa é de que no futuro viremos a acreditar que o conhecimento nas áreas que podem parecer hoje se enquadrarem nas capacidades (robustez, generalização fora da amostra, segurança, outras formas de interpretabilidade, modularidade, interação humano-IA, aprendizado contínuo, motivações intrínsecas) são um componente crítico do nosso portfólio da segurança da IGA. Pelo menos, pode ser útil ter mais trabalho de "pós-doutorado" em todo o espaço do aprendizado de máquinas para então trazer conhecimentos e habilidades às oportunidades mais valiosas na área da segurança da IA.
3. Desenvolver uma carreira em outras áreas que podem ser codificadas como "capacidades" poderia resultar em indivíduos se destacando em vários campos do aprendizado de máquinas e exercendo funções importantes e de influência em organizações de IA. Acredito que é muito vantajoso que a comunidade de pessoas que se preocupam com os riscos da IA tenham uma compreensão ampla do campo do aprendizado de máquinas e das organizações de IA e ampla habilidade de criar normas. Na minha opinião, grande parte dos benefícios de "pesquisadores de segurança da IA" não advém do trabalho que eles realizam mas de sua influência normativa e organizacional na área da ciência do aprendizado de máquinas e nas organizações em que trabalham. Minha expectativa é de que os conhecimentos críticos sobre a segurança terão que ser absorvidos e implementados em campos fora da área de

segurança e portanto é vantajoso ter indivíduos cientes das questões de segurança nesses campos. Dada essa posição, faz sentido diversificar as especialidades seguidas por indivíduos que se preocupam com os riscos da IA e atribuem um peso maior às direções particularmente empolgantes cientificamente ou vantajosas às organizações de IA com o potencial de construir sistemas de IA poderosos.

4. O trabalho com as capacidades já é altamente incentivado ao ritmo de bilhões de dólares e será ainda mais no futuro, então acho que, na margem, os indivídos motivados a combater os riscos da IA que estão trabalhando nesse espaço não aumentariam muito as capacidades.  Para tentar quantificar isso, cerca de 6.000 autores compareceram à NeurIPS em 2021. Aumentar esse número em 1 representa um aumento de 1/6.000. Em contraste, acho que os benefícios à segurança, mencionados acima, de um indivíduo aprender com sua atuação em outros campos, ser potencialmente um líder numa nova área crítica da segurança da IA e estar em posição potencialmente melhor de influenciar as normas e decisões de uma organização provavelmente seriam muito maiores. (Uma crença relevante em meu modo de pensar é que não acredito que encurtar as escalas de tempo *hoje* nos custa tanta segurança assim, em comparação a nos colocarmos numa posicão melhor mais próxima ao período crítico.) Note que este ultimo argumento não se aplica aos grandes atores tais como laboratórios ou patrocinadores importantes.

**Especialista n.o 9: No atual ritmo de progresso com as capacidades da IA em comparação ao nosso progresso com o alinhamento…**

No atual ritmo de progresso com as capacidades da IA em comparação ao nosso progresso com o alinhamento, é improvável que a questão do alinhamento será resolvida a tempo antes da

implementação da primeira IAG. Se você acredita que o alinhamento é improvável logo de início, essa é uma situação lamentável.

*Dada a situação atual, qualquer desaceleração marginal das melhorias das capacidades e qualquer aceleração marginal do trabalho com o alinhamento é importante se pretendermos resolver esse problema a tempo.*

Por esse motivo, os indivíduos preocupados com a segurança da IA deveriam usar de muito cuidado antes de decidir trabalharem com as capacidades e considerar seriamente a possibilidade de trabalhar com o alinhamento e com a segurança da IA diretamente sempre que possível. Este é o caso principalmente porque o campo da IA é relativamente pequeno e tem uma concentração extrema de talento: os maiores pesquisadores e engenheiros de aprendizado de máquinas são unicamente responsáveis por grandes fatias do progresso total.

Portanto, é particularmente importante que pessoas de grande talento escolham muito bem a área em que vão trabalhar: cada pessoa de talento escolhendo trabalhar com a segurança da IA e não com as capacidades causa o dobro do impacto, simultaneamente ganhando mais tempo antes da IAG e ao mesmo tempo acelerando o trabalho com o alinhamento.

Uma coisa crucial a considerar é não apenas a organização mas também a equipe e o tipo de trabalho. Algumas organizações que são criticadas muitas vezes por seu trabalho com as capacidades na comunidade têm equipes que sinceramente se preocupam com o alinhamento, e um trabalho aí provavelmente ajuda. *Do lado oposto, algumas organizações que expressam abertamente seu foco na segurança têm equipes grandes focando na aceleração das capacidades, de forma pública ou privada, e trabalhar dentro dessas equipes é provavelmente prejudicial.*

*A relação entre o trabalho com capacidades e o trabalho vantajoso com o alinhamento não é binária, e grande parte do trabalho com o alinhamento que é mais promissor também tem consequências na área de capacidades, mas o contrário raramente é o caso, e apenas acidentalmente.*

Algumas organizações e indivíduos, entre eles alguns intimamente afiliados ao Altruísmo Eficaz (AE), são de opinião que acelerar o progresso agora é bom porque o problema do alinhamento é essencialmente um problema empírico de engenharia, e modelos mais avançados nos permitiriam conduzir pesquisas empíricas melhores sobre como controlar a IAG.

Uma outra opinião comum é que acelerar o progresso para atores "amigáveis," tais como aqueles que dizem se importar mais com a segurança, têm laços com o AE e localizam-se em países não-autoritários, é necessário, pois preferiríamos que os atores mais conscientes da segurança chegassem à IAG primeiro.

*Aqueles que agem com base nessas opiniões estão sendo extremamente irresponsáveis, e os indivíduos que pretendem trabalhar com a IA deveriam se opor a tais argumentos usados como desculpa para a aceleração das capacidades.*

**Especialista no. 10: Penso que a comunidade do AE parece, em geral, focar demasiadamente no mal da aceleração das capacidades…**

Penso que a comunidade do AE parece, em geral, focar demasiadamente no mal da aceleração das capacidades. Acho mesmo que acelerar as capacidades é mau (todo o resto permanecendo igual), mas o impacto marginal de um engenheiro ou pesquisador é muito pequeno e não parece difícil contrabalançá-lo com os benefícios, entre eles empoderar uma organização a conduzir pesquisas melhores sobre a segurança, ser mais influente, etc.;

adquirir capital de carreira de vários tipos (compreensão da IA, contatos profissionais, realizações, etc.).

Contudo, se você se encontra nessa categoria, eu faria um esforço extra em:

- Ser um funcionário que presta atenção às ações da empresa onde está trabalhando, pede a outras pessoas que lhe ajude a entender as razões por trás das ações e se expresse quando não estiver contente ou à vontade. Penso que você deveria passar 95% ou mais do seu tempo focado em fazer um bom trabalho, e críticas têm mais poder vindo de um funcionário altamente produtivo (se você não estiver se saindo muito bem no trabalho, eu focaria exclusivamente nisso, e/ou me demitiria, em vez de gastar tempo/energia em debates sobre a estratégia e tomadas de decisão da empresa). Mas acho que os 5% restantes podem ser importantes — os funcionários são parte da "consciência" de uma organização.
- Evite estar numa situação financeira ou psicológica em que é claramente difícil para você trocar de função e trabalhar em algo mais exclusivamente focado em fazer o bem; pergunte-se constantemente se você seria capaz de fazer essa mudança e se você está tomando decisões que poderiam tornar esse tipo de mudança mais difícil de realizar no futuro.

**Especialista no. 11: Na minha expectativa, o determinante dos riscos da IA que mais afeta os humanos é…**

Na minha expectativa, o determinante dos riscos da IA que mais afeta os humanos é o grau com que os primeiros 2-5 grandes laboratórios de IA que criarem uma IA transformadora forem capazes de interagir entre si e com o público de boa-fé, a ponto de entrarem num acordo e fazerem cumprir normas impedindo a formação de muitos laboratórios menores que poderiam fazer algo desavisado ou precipitado com sua tecnologia (inclusive tentativas desavisadas ou

precipitadas de reduzir os riscos-x). Para mim, isto nos conduz aos três pontos seguintes:

1. Trabalhar com as capacidades da IA num grande laboratório, de uma maneira que promove relações de boa-fé com esse laboratório, com outros e com o público, é provavelmente líquido positivo na minha opinião. Ressalva: Se você tiver uma ideia genial que permite a criação da IAG 6 meses antes da época em que ela teria sido criada sem a sua intervenção, então é provavelmente líquido negativo divulgar essa conclusão com o seu empregador, mas também você teria alguma liberdade de decidir se deve divulgar isso ou não, de forma que ter um emprego voltado para as capacidades da IA num dos (5) principais laboratórios provavelmente valeria a pena se você puder contribuir de maneira significativa a relações de boa-fé entre esse laboratório, outros laboratórios e o público. Eu consideraria uma "contribuição positiva significativa" algo como "sem engodo, fazer com que o Grande Laboratório A baixe em 1% sua probabilidade subjetiva de que o Grande Laboratório B vai desertar contra o Grande Laboratório A se o Grande Laboratório B alcançar a IA transformadora primeiro." Vamos chamar isso de uma "contribuição de 1% à boa-fé." Penso que uma contrubuição de 0,1% à boa-fé poderia ser pequena demais para justificar o trabalho com as capacidades, e uma contribuição de 10% à boa-fé é mais do que suficiente.
2. Se você acha que sua habilidade de servir como modelo em situações sociais não é adequada a determinar se você está fazendo uma contribuição à qualidade de boa-fé das relações entre os (5) maiores laboratórios de IA e o público, minha sugestão é que você provavelmente não deveria tentar trabalhar com a pesquisa sobre as capacidades da IA em nenhum contexto porque você não vai estar em boas condições de avaliar se o laboratório em que você trabalha está criando as

capacidades de uma maneira que aumente a boa-fé entre os laboratórios de IA e outros a seu redor.

3. Trabalhar com as capacidades rudimentares da IA num laboratório pequeno é provavelmente bom contanto que você não esteja impulsionando a tecnologia de ponta e contanto que você não esteja participando de uma grande deserção contra os grandes laboratórios que estão tentando prevenir a propagação de tecnologias prejudiciais à sociedade. Por exemplo, acho que você não deveria tentar reproduzir GPT-4 e lançá-lo ou implementá-lo em maneiras que acabam por dispensar todo o trabalho árduo que a equipe da OpenAI terá realizado para assegurar que sua versão do modelo esteja sendo usada eticamente.

**Converse individualmente com a nossa equipe**

Se você está pensando em exercer uma função que poderia melhorar as capacidades da IA ou está considerando essa questão de forma geral com relação à sua carreira, nossa equipe de consultores pode lhe oferecer aconselhamento pessoal. (É grátis.) Estamos empolgados em apoiar alguém que deseja focar sua carreira na redução dos riscos impostos pela IA. Nossa equipe pode ajudar você a comparar opções, entrar em contato com outras pessoas que estão trabalhando com esta questão e talvez até ajudar você a encontrar empregos ou oportunidades de patrocínio.

FALE COM A NOSSA EQUIPE