Global Coalition for Tech Justice

Proposta para a Cúpula de Ação sobre Inteligência Artificial

10 e 11 de fevereiro de 2025, França

Grupo de Trabalho sobre
IA e Integridade da Informação
Coalizão Global para Justiça Tecnológica (GCTJ)
Novembro de 2024

Eixos temáticos:
Confiança na IA, Governança Global de IA e IA de Interesse Público

As seguintes propostas foram elaboradas pelo Grupo de Trabalho de IA da Coalizão Global para Justiça Tecnológica.

A Coalizão Global para Justiça Tecnológica é composta por 250 organizações e especialistas em 55 países. O principal objetivo da Coalizão é garantir que as empresas de Big Tech cumpram seu papel na proteção da democracia e dos direitos humanos no mundo inteiro, especialmente na maioria global, onde as empresas têm sido negligentes ao lidar com os impactos gerados por seus modelos de negócios e tecnologias emergentes.

Os co-signatários do estão listados ao fim do documento, mas a proposta permanecerá aberta para assinatura até fevereiro de 2025.

A Digital Action é a organizadora da Global Coalition for Tech Justice. Desde 2019, a Digital Action vem mobilizando uma rede global de parceiros para exigir melhores padrões dos governos e corporações responsáveis por nossos ambientes digitais.

Índice

I. Introdução

Do planejamento à implementação, da elaboração de políticas à prestação de contas e responsabilização, existe uma desigualdade global profunda no cerne da Inteligência Artificial (IA), que está moldando cada vez mais o futuro da integridade da informação, da democracia e dos direitos humanos.[1] Essa desigualdade representa um desafio fundamental para a justiça global e a governança democrática na era digital. À medida que os sistemas de IA se tornam mais profundamente incorporados em instituições públicas e infraestruturas[2] essenciais - de saúde e educação à sistemas judiciais e serviços públicos -, esse desequilíbrio de poder ameaça exacerbar as disparidades globais existentes e criar novas formas de dependência tecnológica que minam a autonomia nacional e a autodeterminação. A rápida aceleração do desenvolvimento da IA, concentrada em alguns poucos centros globais de poder,[3] apresenta o risco de consolidar essas desigualdades nos alicerces do nosso futuro digital compartilhado.

A IA é principalmente desenvolvida no Norte Global ou na China e implementada em todas as regiões do mundo, com mínima consideração pelos contextos ou consequências locais. Enquanto os prejuízos à Maioria Global são sistematicamente ignorados, as capacidades e a expertise necessárias para a elaboração de políticas sobre IA que respeitem os direitos permanecem insuficientes nessas regiões. Estes prejuízos vão desde o viés algorítmico[4] até a deslocação dos sistemas locais de tomada de decisão. Os impactos se manifestam de várias maneiras: sistemas de IA que falham em reconhecer idiomas locais e nuances culturais, sistemas de tomada de decisão automatizados que são treinados com dados ocidentais que fazem determinações inadequadas em contextos do Sul Global, e sistemas de moderação de conteúdo orientados por IA que, inadvertidamente, suprimem linguagem política legítima.[5] A falta de conhecimento especializado e de recursos locais para identificar e lidar com esses prejuízos, juntamente com a exclusão sistemática, por parte das empresas, do conhecimento e da experiência do Sul Global no desenvolvimento e na implantação da IA, cria um ciclo vicioso de dependência tecnológica e marginalização.

Há uma profunda assimetria na capacidade de aplicação da lei e no alcance jurisdicional, de modo que a maioria dos países da Maioria Global acaba não partcipando dos processos de tomada de decisão que moldam efetivamente os sistemas de IA implantados em seus espaços de

---

[1] United Nations, 'Urgent Action Needed over Artificial Intelligence Risks to Human Rights' (UN News, 17 September 2021) https://news.un.org/en/story/2021/09/1099972 accessed 5 May 2024.

[2] General Purpose Technologies (GPTs) 'are technologies that, throughout history, have changed the entire economy and, therefore, have the potential to implement drastic changes in society with an impact on pre-existing economic and social structures'. André Guidetti, Artificial Intelligence as General Purpose Technology: An Empirical and Applied Analysis of its Perception (Master's Thesis, Università della Valle d'Aosta - Université de la Vallée d'Aoste 2020), p.1 https://univda.unitesi.cineca.it/bitstream/20.500.14084/428/1/ETI_104_Guidetti_André.pdf accessed 7 October 2024.

[3] Anu Bradford, 'The Race to Regulate Artificial Intelligence' (Foreign Affairs, 27 June 2023) https://www.foreignaffairs.com/united-states/race-regulate-artificial-intelligence-sam-altman-anu-bradford accessed 25 October 2024

[4] United Nations, 'Impact of New Technologies on the Promotion and Protection of Human Rights in the Context of Assemblies, Including Peaceful Protests' (2020) https://undocs.org/Home/Mobile?FinalSymbol=A%2FHRC%2F44%2F24&Language=E&DeviceType=Desktop&LangRequested=False accessed 8 October 2024.

[5] Frederik Zuiderveen Borgesius, 'Discrimination, Artificial Intelligence, and Algorithmic Decision-Making' (Council of Europe, 2018) https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73 accessed 8 October 2024.

informação, mesmo que os regulamentem. Esse desequilíbrio de poder prejudica a soberania nacional e a governança democrática no âmbito digital. Até quando os países desenvolvem regulamentações abrangentes de IA, eles enfrentam desafios significativos para aplicar essas regras contra poderosas empresas multinacionais de tecnologia. A natureza transnacional dos sistemas de IA, combinada com a concentração de conhecimento técnico e jurídico no Norte Global, cria uma situação em que as nações da Maioria Global geralmente precisam aceitar quaisquer sistemas e políticas de IA que lhes sejam impostos, independentemente das leis ou normas sociais locais.

E no decorrer destes processos, cidadãos e a sociedade civil organizada acabam sendo marginalizados ou excluídos em partes do processo: desenvolvimento, implementação, prestação de contas e elaboração de políticas. Eles precisam de acesso, bem como de suporte contínuo para desenvolver capacidade e expertise visando uma inclusão significativa. Essa exclusão perpetua um ciclo de dependência tecnológica e déficit democrático. As organizações da sociedade civil, que tradicionalmente desempenham papéis cruciais na proteção do interesse público e na promoção da participação democrática, muitas vezes não têm o conhecimento técnico e os recursos necessários para se envolver com a governança da IA de forma eficaz. A complexidade dos sistemas de IA e a opacidade deliberada em seu desenvolvimento e implantação criam barreiras substanciais para uma participação pública significativa. Essa exclusão sistemática das vozes dos cidadãos significa que os sistemas de IA são desenvolvidos sem a contribuição crucial das comunidades que eles mais afetarão.

É exatamente por isso que a transparência e a explicabilidade devem ser princípios essenciais no desenvolvimento de uma IA ética. Adicionalmente, as pessoas têm o direito fundamental de entender como os sistemas de IA afetam suas vidas e de receber explicações claras sobre as decisões automatizadas.[6] A transparência atende a duas funções cruciais: permite que o público entenda como os sistemas de IA funcionam e, mais importante, fornece a base necessária para responsabilizar os desenvolvedores e as plataformas pelos impactos de suas tecnologias. Sem essa transparência, a participação efetiva dos cidadãos e a supervisão permanecem impossíveis, reforçando ainda mais o desequilíbrio de poder entre os desenvolvedores de IA e as comunidades afetadas por seus sistemas.

Uma das consequências da desigualdade global, conforme descrito, é a incorporação da desigualdade racial no desenvolvimento,[7] implantação,[8] acesso à responsabilização e na elaboração de políticas sobre IA. Esse viés sistêmico[9] agrava as injustiças sociais existentes e

---

[6] Gabriela Arriagada Bruneau, Los sesgos del algoritmo: La importancia de diseñar una inteligencia artificial ética e inclusiva [The Biases of the Algorithm: The Importance of Designing an Ethical and Inclusive Artificial Intelligence] (La Pollera, 2024) https://lapollera.cl/libros/sesgos-algoritmo-ia-etica/ accessed 28 October 2024.

[7] 'Every dataset used to train machine learning systems, whether in the context of supervised or unsupervised machine learning, whether seen to be technically biased or not, contains a worldview'. Kate Crawford, *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (Yale University Press 2021) 139

[8] Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' in *Proceedings of Machine Learning Research*, Conference on Fairness, Accountability, and Transparency (2018) 81:1–15 http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf accessed 21 October 2024.

[9] Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown 2016) 10

ameaça consolidar padrões históricos de discriminação. A falta de diversidade[10] nas equipes de desenvolvimento de IA, nos dados de treinamento e nos processos de teste acarretam em sistemas que não apenas deixam de abordar as desigualdades raciais existentes, mas também as reforçam ativamente. Desde sistemas de reconhecimento facial que têm um mau desempenho com tons de pele mais escuros[11] até modelos de linguagem que perpetuam estereótipos nocivos, as implicações raciais das atuais práticas de desenvolvimento de IA são profundas e de longo alcance. A ausência de mecanismos eficazes de responsabilização e análises de risco significa que esses vieses muitas vezes não são detectados e não são abordados até que um dano significativo já tenha ocorrido.

No mundo digital de hoje, a integridade da informação é fundamentalmente uma questão de capacitar as pessoas a assumir o controle de suas informações - o que elas acessam, o que consomem, como as informações são apresentadas e como elas as avaliam. No entanto, estamos enfrentando um desafio crítico: um punhado de grandes empresas de tecnologia (Big Tech) detém atualmente um poder sem precedentes sobre nosso ecossistema de informações, determinando o que bilhões de pessoas veem e como veem. Essa concentração de poder não é apenas uma questão comercial - ela influencia diretamente e, muitas vezes, prejudica o debate democrático.

A integridade da informação requer três elementos essenciais: transparência na forma como as informações são selecionadas e distribuídas, responsabilização daqueles que controlam esses sistemas e uma rica pluralidade de fontes de informações confiáveis. Embora os Princípios Globais[12] para a Integridade da Informação da ONU forneçam uma estrutura importante - enfatizando a confiança da sociedade, incentivos saudáveis, capacitação do público, mídia independente e transparência na pesquisa - sua perspectiva centrada no Norte Global exige uma análise crítica. Do ponto de vista do Sul Global,[13] devemos abordar desafios estruturais mais profundos: soberania digital em face da resistência das plataformas de redes sociais a acatar regras de governança local, a necessidade de promover o jornalismo como uma prática ética em vez de apenas "mídia independente", e a reconstrução de espaços sociais comuns em vez de simplesmente criar resistência contra ameaças externas. Precisamos ir além de um mundo em que as plataformas de Big Tech e seus algoritmos dominam a curadoria de informações, reconhecendo, ao mesmo tempo, que o empoderamento público significativo exige a abordagem das desigualdades sociais, raciais e de gênero fundamentais. Os Estados democráticos devem assumir papéis ativos na prevenção da concentração de mercado e na garantia de acesso equitativo a dados e oportunidades de pesquisa, especialmente para pesquisadores do Sul Global

[10] Forum on Information and Democracy, 'AI as a Public Good: Ensuring Democratic Control of AI in the Information Space' (February 2024) https://informationdemocracy.org/wp-content/uploads/2024/03/ID-AI-as-a-Public-Good-Feb-2024.pdf accessed 10 October 2024 p.24

[11] Morgan Meaker, 'This Student Is Taking On "Biased" Exam Software' (WIRED, 5 April 2023) https://www.wired.com/story/student-exam-software-bias-proctorio/

[12] United Nations, 'Global Principles for Information Integrity: Recommendations for Multi-stakeholder Action' (2023) https://www.un.org/sites/un2.un.org/files/un-global-principles-for-information-integrity-en.pdf accessed 23 October 2024.

[13] Nina Santos, 'Five Brazilian Principles for the Integrity of the Information Ecosystem' (Tech Policy Press, 2 November 2023) https://www.techpolicy.press/five-brazilian-principles-for-the-integrity-of-the-information-ecosystem/ accessed 11 November 2024

que atualmente enfrentam barreiras sistêmicas. Essa transformação exige não apenas inovação técnica e supervisão democrática, mas também o compromisso de abordar as desigualdades estruturais que moldam nosso ecossistema de informações.

O Órgão Consultivo de Alto Nível sobre Inteligência Artificial estabeleceu princípios e recomendações cruciais para a governança global da IA, enfatizando que esses não podem ser efetivamente realizados sem abordar desigualdades fundamentais, especialmente entre o Norte e o Sul Global. Embora a visão do Órgão Consultivo de governança inclusiva e baseada em direitos forneça uma base essencial, a transformação desses princípios em realidade requer mecanismos específicos e acionáveis para redistribuir o poder no ecossistema de IA. Nossa proposta de uma "AI Commons" oferece um caminho prático para implementar esses princípios, garantindo que vozes historicamente marginalizadas possam participar de forma significativa na definição do desenvolvimento e da implantação dessas tecnologias. Essa abordagem trata diretamente das preocupações do Órgão Consultivo sobre as lacunas de representação e os desafios de implementação das atuais iniciativas internacionais de governança de IA, fornecendo soluções tangíveis para a construção de uma verdadeira equidade global na governança de IA.

Nossa proposta busca abordar a desigualdade global e seus inúmeros componentes, principalmente a desigualdade geográfica, racial e social, por meio de uma intervenção coordenada e sistemática em múltiplos níveis. Essa abordagem abrangente reconhece que o enfrentamento à desigualdade da IA exige ações simultâneas nos domínios técnico, social e político. Ela exige a criação de novas instituições e estruturas que possam redistribuir efetivamente o poder no ecossistema de IA e, ao mesmo tempo, desenvolver a capacidade local para uma participação significativa na governança da IA. Ao abordar a natureza interconectada dessas desigualdades, nosso objetivo é criar uma mudança sustentável que empodere as comunidades a moldar os sistemas de IA que afetam suas vidas.

Não se trata apenas de criar novas regulamentações - trata-se de redistribuir fundamentalmente o poder no espaço de informações digitais e garantir que a tecnologia sirva à democracia e aos direitos humanos ao invés de prejudicá-los. Para esses fins, estamos propondo os seguintes resultados implementáveis para a Cúpula Global de IA em fevereiro de 2025, projetados para criar impacto imediato e, ao mesmo tempo, desenvolver capacidade de longo prazo para a governança democrática dos sistemas de IA.

II. AI Commons: Democratizando a IA por meio do poder do cidadão global

Propomos o lançamento de uma iniciativa "AI Commons" com quatro pilares de implementação. Imagine uma rede global em que os cidadãos, especialmente aqueles que nunca tiveram um lugar à mesa antes, possam ajudar a moldar como a IA é desenvolvida e utilizada. Por meio de quatro programas interconectados - centros de treinamento em todo o Sul Global, conselhos de cidadãos que analisam os projetos e políticas de IA, laboratórios onde as pessoas podem experimentar novas políticas de IA e um sistema de supervisão abrangente que garante a prestação de contas e a responsabilização - colocamos o poder da IA nas mãos de todos. Cada pilar desempenha um

papel fundamental: a Rede de Laboratórios de Equidade de IA[14] desenvolve a capacidade, o Conselho Cidadão de Desenvolvimento (CDC)[15] permite a contribuição direta da comunidade, o Laboratório de Inovação em Políticas de IA (APIL)[16] permite a experimentação segura e o Sistema de supervisão de múltiplas partes interessadas (MOS)[17] garante que todo o processo permaneça responsável e transparente. Não se trata apenas de tornar a IA mais justa; trata-se de garantir que ela funcione para todos nós, não apenas para alguns poucos selecionados. Essa não é uma visão distante para o futuro - ela pode acontecer agora, fornecendo ferramentas reais para que as pessoas possam garantir que a IA ajude, e não atrapalhe, nossas sociedades democráticas.

A. Rede de Laboratórios de Equidade de IA

A Rede de Laboratórios de Equidade de IA pode ser vista como uma escola global para os futuros formadores de políticas sobre IA. Ainda assim, com uma diferença: ela foi projetada especificamente para dar poder às vozes de comunidades que tradicionalmente não tem sido consideradas no desenvolvimento da tecnologia. Por meio de centros físicos espalhados pela Maioria Global e de uma plataforma on-line robusta, ela está criando espaços onde as pessoas podem ter experiência prática com sistemas de IA e, ao mesmo tempo, aprender a orientar seu desenvolvimento na direção certa.

O que torna essa rede especial é sua abordagem abrangente. Não se trata apenas de treinamento técnico - os participantes passam por um programa de bolsas de um ano em que aprendem de tudo, desde a auditoria de sistemas de IA para garantir a justiça até a elaboração de políticas que protejam os interesses de suas comunidades. Até 2026, a rede pretende treinar 1.000 novos líderes de políticas de IA da África, Ásia, Oriente Médio e América Latina que entendam os aspectos técnicos e sociais da IA, criando uma força poderosa para mudanças positivas no cenário global da IA.

Modelo de financiamento e parcerias

A sustentabilidade financeira da Rede de Laboratórios de Equidade de IA se baseia em um cuidadosamente elaborado modelo de financiamento de participação de multissetorial. Em vez de depender de uma única fonte de financiamento, criamos uma abordagem equilibrada que distribui recursos e responsabilidades entre diferentes setores. A participação do governo dos países anfitriões oferece suporte e legitimidade de infraestrutura essenciais. Seu investimento demonstra um compromisso com o desenvolvimento de conhecimento especializado em políticas locais de IA e garante que o programa se alinhe às metas nacionais de desenvolvimento.

As parcerias com o setor privado trazem mais do que apenas apoio financeiro. As principais empresas de tecnologia fornecem recursos técnicos essenciais, oportunidades de orientação e estudos de caso do mundo real. No entanto, estruturamos essas parcerias visando manter a

[14] Originalmente em inglês, Equity Lab Network
[15] Originalmente em inglês, Citizens' Design Council
[16] Originalmente em inglês, Policy Innovation Lab
[17] Originalmente em inglês, Multi-stakeholder Oversight System

independência da rede e a capacidade de avaliar criticamente as tecnologias de IA e seus impactos.

As instituições internacionais são fundamentais para garantir a relevância e a sustentabilidade global do programa. Seu envolvimento ajuda a manter os altos padrões e facilita o compartilhamento de conhecimento entre as regiões. O modelo de financiamento inclui mecanismos para a sustentabilidade de longo prazo, incluindo um fundo patrimonial e atividades geradoras de receita que apoiam o crescimento da rede e, ao mesmo tempo, mantêm sua missão principal.

Currículo e estrutura de treinamento

O currículo da Rede de Laboratórios de Equidade de IA apresenta uma abordagem que busca preencher a lacuna entre o conhecimento técnico e a compreensão da política na governança da IA. Em sua essência, o programa reconhece que líderes eficazes em políticas de IA precisam de um entendimento abrangente que abranja as dimensões técnica e social. Para conseguir isso, desenvolvemos um sofisticado sistema de curso de trilha dupla que se adapta ao histórico dos participantes, garantindo que todos obtenham um conjunto de habilidades multidisciplinares.

A trilha "Sistemas de IA & Elaboração de Políticas" para profissionais da área jurídica e de política começa com a desmistificação da tecnologia de IA. Os participantes começam com experiência prática em conceitos básicos de programação e *machine learning*, indo além da compreensão teórica para a aplicação prática. Por meio de laboratórios interativos e projetos do mundo real, eles aprenderão a avaliar os sistemas de IA de forma crítica, a entender suas limitações e a avaliar seus impactos sociais. Ao final do programa, esses participantes poderão se comunicar de forma eficaz com equipes técnicas e tomar decisões políticas informadas com base em um entendimento técnico genuíno.

Já os profissionais técnicos que ingressam no programa seguem a trilha "Integração de políticas, direitos humanos e ética", transformando sua experiência técnica em conhecimento relevante para as diretrizes relativas ao uso de IA. Esse curso enfatiza o cenário regulatório, as estruturas internacionais de direitos digitais e as formas diferenciadas como a IA afeta diferentes comunidades. Os participantes aprendem a traduzir seu conhecimento técnico em recomendações de diretrizes, considerando diversos contextos culturais e necessidades sociais.

Ambas as trilhas convergem em um currículo básico comum que desenvolve habilidades cruciais de liderança e incidência política. Essa experiência compartilhada cria uma rede poderosa de profissionais que podem superar a divisão tradicional entre os domínios técnico e político.

Seleção e distribuição regional

O processo de seleção de participantes para a Rede de Laboratórios de Equidade de IA foi pensado para criar uma comunidade diversificada, de alto impacto, de futuros líderes para políticas em relação ao uso de IA. A partir da ideia que diferentes regiões enfrentam desafios e oportunidades únicas no desenvolvimento da IA, o nosso objetivo é estabelecer um sistema de cotas equilibrado que garanta a representação de toda a Maioria Global. Não se trata apenas de números - mas de criar um diálogo frutífero entre diferentes perspectivas e experiências.

Nossos critérios de seleção vão além das métricas tradicionais. Embora a experiência profissional seja importante, pretendemos valorizar também os candidatos que demonstram um profundo entendimento de seus contextos regionais e apresentam potencial para catalisar mudanças em suas comunidades. Para isso, buscamos pessoas que possam fazer a ponte entre os desenvolvimentos globais de IA e as necessidades locais, considerando fatores como seu envolvimento com iniciativas comunitárias e sua capacidade de navegar em complexas relações com as partes interessadas.

A estrutura física do centro (*hub*) é fundamental para a nossa visão. Cada *hub* - seja em Nairóbi, Jacarta ou São Paulo - funcionará como um centro de excelência para sua região, adaptado aos contextos locais e mantendo os padrões globais. Esses *hubs* não são apenas centros de treinamento, mas também incubadoras para inovação de políticas regionais de IA pensados para promover a colaboração entre os participantes e as partes interessadas locais.

Implementação e estrutura de governança

A estratégia de implementação reflete nosso compromisso com a construção de uma instituição duradoura que possa se adaptar e crescer. Nossa estrutura de governança combina supervisão global com autonomia regional, garantindo que os programas permaneçam relevantes para os contextos locais ao mesmo tempo que mantêm os padrões internacionais. O Conselho Consultivo Internacional desempenha um papel fundamental na direção estratégica, trazendo diversas perspectivas de especialistas técnicos, especialistas em políticas e líderes da sociedade civil.

A garantia de qualidade é incorporada em todos os aspectos do programa. Revisões regulares do currículo, mecanismos de *feedback* dos participantes e auditorias externas são algumas das medidas que visam garantir que a rede continue a atender à evolução das necessidades do desenvolvimento de diretrizes em relação ao uso de IA. A avaliação do impacto vai além das métricas tradicionais para avaliar como os bolsistas influenciam as políticas de IA em suas regiões e criam mudanças positivas em suas comunidades.

O cronograma de implementação é ambicioso, mas realista. Começar com regiões-piloto nos permite aprimorar nossa abordagem antes do escalonamento. Nossa meta de treinar 1.000 líderes de políticas de IA até 2026 não se trata apenas de números - trata-se de criar uma massa crítica de conhecimento em regiões que historicamente têm sido sub-representadas nas discussões globais de governança de IA.

B. Conselho Cidadão de Desenvolvimento (CDC)

O Conselho Cidadão de Desenvolvimento (CDC) está revolucionando a forma como a IA é desenvolvida, trazendo pessoas comuns para o processo de desenvolvimento. Com os principais centros regionais localizados na África, na América Latina, na Ásia e no Oriente Médio, ele está criando uma estrutura em que as comunidades - especialmente aquelas que geralmente são negligenciadas no desenvolvimento de tecnologia - têm uma opinião real sobre como os sistemas de IA são criados e implantados em suas regiões.

Não se trata apenas de *feedback* após o fato ocorrido - o CDC envolve comunidades em todas as etapas do desenvolvimento da IA. Antes da criação de qualquer sistema , seus membros

avaliarão o impacto cultural e as necessidades da comunidade em torno de uma iniciativa. Durante o desenvolvimento, eles testarão protótipos e monitoram os impactos. E após a implementação, eles verificarão continuamente como esses sistemas afetam a vida das pessoas reais. Trata-se de garantir que a IA funcione para todos, não apenas para os poucos conhecedores de tecnologia.

C. Laboratório de Inovação em Políticas de IA (APIL)

Imagine um espaço onde diversas vozes - formuladores de políticas, ativistas da sociedade civil, comunidades afetadas, academia, defensores de direitos humanos e entidades privadas - possam ver e experimentar de forma colaborativa como as políticas de IA afetam os direitos humanos fundamentais e a vida cotidiana das pessoas antes da implementação. É isso que o Laboratório de Inovação em Políticas de IA (APIL) oferece. Suas ferramentas de visualização de ponta e espaços de simulação reúnem tecnólogos, especialistas em direitos humanos, líderes comunitários, representantes de empresas e formuladores de políticas para transformar ideias abstratas de políticas em cenários tangíveis que demonstram os impactos sobre a privacidade, a liberdade de expressão, a não discriminação e outros direitos humanos essenciais no mundo real.

O laboratório combinará ferramentas de alta tecnologia com a elaboração de políticas centradas em direitos humanos por meio de uma abordagem inovadora de participação de multissetorial. Em suas estações de trabalho colaborativas e salas de políticas públicas de realidade virtual, líderes indígenas trabalharão ao lado de defensores de direitos digitais; organizações de base fazem parcerias com funcionários do governo e especialistas acadêmicos unirão forças com representantes da juventude para testar como diferentes políticas públicas de IA podem afetar comunidades vulneráveis e liberdades fundamentais. Por meio de simulações imersivas, esse grupo diversificado poderá experimentar em primeira mão como as decisões podem afetar o acesso à informação, a proteção da privacidade ou ampliar a discriminação. Durante os esforços será dada atenção especial aos impactos interseccionais, com as comunidades afetadas liderando a conversa sobre como as políticas podem impactar de forma diferente as pessoas com base em seu gênero, etnia, status econômico ou localização geográfica. O objetivo é ter área de criatividade para experimentar políticas com consciência e sabedoria coletiva, onde as ideias podem ser testadas com segurança e aprimoradas por meio de várias perspectivas para garantir que protejam e fortaleçam os direitos humanos antes de afetar milhões de vidas. A abordagem participativa do laboratório garante que as políticas não sejam criadas apenas para as comunidades, mas com as comunidades, ajudando a evitar consequências indesejadas que possam comprometer a dignidade humana ou exacerbar as desigualdades existentes.

D. Sistema de Supervisão Multissetorial (MOS)

O cenário atual da IA apresenta um paradoxo crítico. Enquanto empresas como OpenAI, Meta e gigantes regionais da tecnologia implantam rapidamente sistemas de IA em nível global, os mecanismos de supervisão permanecem fragmentados e, muitas vezes, ineficazes. Testemunhamos contradições gritantes: O controle rigoroso da China sobre o acesso ao ChatGPT versus a abordagem amplamente autorregulatória dos EUA, ou as regulamentações centradas em direitos da Europa em comparação com as estruturas emergentes das regiões em desenvolvimento (limitadas pelos desafios de equidade descritos acima). Essas disparidades destacam a necessidade urgente de um sistema de supervisão equilibrado e adaptável para

conciliar essas abordagens e, ao mesmo tempo, priorizar os direitos humanos e a integridade das informações.

A composição deste órgão deve ser cuidadosamente equilibrada e será multissetorial desde a sua concepção. Especialistas técnicos trabalham ao lado de defensores de direitos humanos; especialistas jurídicos colaboram com jornalistas, e representantes da sociedade civil garantem que as vozes das comunidades permaneçam no centro de todas as decisões. Essa diversidade não se trata apenas de representação - trata-se de reunir as habilidades necessárias para entender as implicações técnicas dos sistemas de IA e seu impacto sobre as comunidades no mundo real.

Este sistema reconhece que a governança da IA enfrenta diferentes desafios em cada região. Em Uganda, onde o governo está revisando sua estratégia de IA, o MOS poderia fornecer uma estrutura para uma supervisão significativa e, ao mesmo tempo, apoiar a inovação local. Em regiões como o Sudeste Asiático, onde a localização de dados e os interesses do Estado desempenham um papel significativo, o sistema poderia oferecer mecanismos flexíveis que respeitem a soberania e, ao mesmo tempo, garantam a proteção dos direitos humanos.

A capacidade deadaptação do sistema,enquanto mantém seus princípios fundamentais, o torna particularmente poderoso. Em regiões onde a autorregulação predomina, ele poderá oferecer mecanismos de supervisão estruturados. Em áreas com forte controle estatal, ele podera oferecer canais para a contribuição da comunidade e a proteção dos direitos. Essa adaptabilidade garante que o sistema permaneça relevante e eficaz em diferentes ambientes regulatórios.

Considere um exemplo prático: Quando uma grande empresa de IA quiser implantar um novo modelo de linguagem na África Ocidental, o órgão de supervisão regional avaliará não apenas as especificações técnicas, mas também as implicações culturais, as preocupações com a privacidade dos dados e os possíveis impactos nos ecossistemas de informações locais. Essa avaliação não se trata de um mero exercício de ticar caixinhas - é uma avaliação abrangente que pode levar a modificações significativas ou até mesmo a restrições de implantação, se necessário.

O MOS garante que o desenvolvimento e a implantação da IA permaneçam transparentes e responsáveis perante todas as comunidades afetadas, tendo como ponto central a proteção dos direitos humanos e a integridade da informação. Ele criará uma estrutura abrangente em que os órgãos regionais de supervisão, compostos por representantes locais, defensores dos direitos humanos, jornalistas, verificadores de fatos, organizações de mídia independentes, membros da sociedade civil e comunidades afetadas, têm poder real para monitorar, avaliar e influenciar como os sistemas de IA afetam os direitos humanos e a integridade da informação em suas regiões.

O MOS supervisionará os sistemas de IA implantados e pré-implantados que afetam significativamente os fluxos de informação e os direitos humanos na sociedade, com foco específico em quatro áreas críticas: (1) Modelos de linguagem de grande escala e sistemas de IA generativa que influenciam o debate público e a criação de informações, (2) sistemas de recomendação e moderação de conteúdo que moldam a distribuição e o acesso às informações, (3) sistemas automatizados de tomada de decisão que afetam os direitos fundamentais e os serviços públicos e (4) tecnologias de vigilância e monitoramento baseadas em IA. Essa supervisão engloba uma avaliação abrangente dos impactos desses sistemas sobre a integridade

das informações, incluindo sua responsabilidade em ampliar ou atenuar a desinformação; seus efeitos sobre o pluralismo da mídia; e sua influência sobre o viés algorítmico na distribuição de informações. Ao mesmo tempo, essa supervisão mantém uma rigorosa responsabilização sobre os direitos humanos por meio de avaliações de impacto obrigatórias, mecanismos de queixa vinculativos e programas de monitoramento liderados pela comunidade. O sistema emprega uma abordagem de via dupla: supervisão proativa por meio de avaliações pré-implantação; e monitoramento contínuo por meio de avaliação pós-implantação, com atenção especial aos sistemas implantados por grandes empresas de tecnologia e entidades governamentais. O processo de supervisão está ancorado em requisitos de documentação transparentes, audiências públicas regulares e mecanismos claros de aplicação, garantindo que o desenvolvimento e a implantação da IA permaneçam responsáveis perante as comunidades afetadas e, ao mesmo tempo, protejam a integridade das informações e os direitos humanos.

O MOS transformará a supervisão de tecnologias de IA por meio de cinco abordagens principais alinhadas aos Princípios da ONU[18] sobre Integridade da Informação:

1. **Construção de confiança:** Estabelecimento de mecanismos de verificação para conteúdo gerado por IA e promoção da transparência em sistemas algorítmicos
2. **Reestruturação de incentivos:** Fazer com que as plataformas transicionem de métricas baseadas em engajamento para métricas de qualidade da informação
3. **Empoderamento do público:** Apoiar a alfabetização digital e criar ferramentas para a supervisão pública dos sistemas de IA
4. **Proteção da mídia:** Proteger a independência e a diversidade jornalística na era da IA
5. **Acesso à pesquisa:** Garantir que os pesquisadores possam estudar de forma significativa o impacto da IA nos ecossistemas de informação

Por meio de sua abordagem integrada centrada em direitos, o MOS garante que os sistemas de IA respeitem os direitos humanos e contribuam para um ambiente de informações saudável, confiável e diversificado. A estrutura de responsabilização e prestação de contas inclui mecanismos de cumprimento que variam de avisos públicos e multas a restrições operacionais para violações graves dos padrões de direitos humanos ou dos princípios de integridade da informação.

- **Um novo modelo para a supervisão da IA**

O MOS representa uma abordagem de supervisão de IA posicionada separadamente entre os observatórios da sociedade civil e os órgãos reguladores formais. Diferentemente dos tribunais que tomam decisões vinculantes ou dos órgãos reguladores que aplicam regras, o MOS funcionará como um observatório transparente cujo objetivo é esclarecer como os sistemas de IA afetam nosso ecossistema de informações e os direitos humanos. Seu poder não está na aplicação da lei, mas em sua capacidade de reunir evidências, revelar padrões e permitir um debate público informado sobre o impacto social da IA. A função de observatório do MOS se alinha aos modelos emergentes de supervisão não regulatória da IA, semelhante ao mecanismo de

[18] United Nations, 'Global Principles for Information Integrity: Recommendations for Multi-stakeholder Action' (2023) https://www.un.org/sites/un2.un.org/files/un-global-principles-for-information-integrity-en.pdf accessed 23 October 2024.

certificação voluntária para IA de interesse público[19] que se mostrou eficaz em outros âmbitos. Conforme destacado na pesquisa[20] do Fórum sobre Informação e Democracia esses mecanismos podem ajudar a resolver as assimetrias de informação entre as empresas provedoras de tecnologias de IA e o público e, ao mesmo tempo, criar incentivos positivos para o desenvolvimento responsável. Assim como os órgãos de certificação bem-sucedidos que mantêm a independência tanto do setor quanto do governo e, ao mesmo tempo, promovem a transparência e a responsabilidade, o MOS servirá como um intermediário confiável que permite o envolvimento significativo das partes interessadas e o escrutínio público. Sua posição como observatório independente permite documentar e analisar os impactos sociais dos sistemas de IA, evitando os riscos de captura regulatória ou influência política que podem afetar órgãos de supervisão mais formais. Essa abordagem permite que o MOS promova a confiança e a legitimidade por meio de avaliação rigorosa e documentação pública, e não por meio de poderes de aplicação da lei.

- **Legitimidade democrática por meio de estrutura e processo**

Em sua essência, a legitimidade do MOS decorre de sua estrutura profundamente democrática. A liderança é rotativa entre representantes de diferentes regiões e grupos de partes interessadas, com limites rígidos de mandatos que impedem o domínio de uma única perspectiva. Um processo de nomeação transparente visa garantir uma representação diversificada, enquanto claras políticas sobre conflito de interesses e fontes de financiamento diversificadas o protegerão o orgão contra a influência por interesses poderosos. Por meio de auditorias independentes regulares de suas próprias operações, o MOS garantirá a mesma transparência que defende nos sistemas de IA.

- **Criando confiabilidade por meio de uma metodologia rigorosa**

As avaliações do MOS ganham credibilidade por meio do rigor metodológico e não da autoridade regulatória. Suas estruturas de avaliação nascem de consultas públicas e passam por revisão por pares, garantindo que reflitam diversas perspectivas e pesquisas atuais. Antes de publicar qualquer descoberta, várias equipes independentes devem verificar os resultados, com a documentação completa de seus métodos tornada pública. Essa abordagem cria percepções confiáveis sobre os impactos sociais dos sistemas de IA e, ao mesmo tempo, mantém limites claros entre observação e determinações.

- **Salvaguardas contra o desvio da missão original e contra a concentração de poder**

Para evitar que o MOS se torne um censor *de facto* ou um judiciário paralelo, seu estatuto proíbe explicitamente os poderes de moderação de conteúdo e limita seu escopo a questões sistêmicas em vez de casos individuais. Revisões externas regulares avaliam a adesão a essas limitações, enquanto relatórios de transparência obrigatórios detalham suas atividades e processos de tomada de decisão. Um sistema robusto de proteção a autores de denúncia incentiva a responsabilidade interna, garantindo que o MOS permaneça fiel à sua missão.

[19] Originalmente em inglês, “voluntary certification mechanism for public interest AI”.
[20] Forum on Information and Democracy, A Voluntary Certification Mechanism for Public Interest AI: Exploring the Design and Specifications of Trustworthy Global Institutions to Govern AI (Research Paper, September 2024).

- **Abordagem colaborativa para supervisão**

Em vez de trabalhar isoladamente, o MOS colabora ativamente com as instituições existentes, respeitando seus distintos papéis. Ele fornece evidências e percepções aos tribunais e órgãos reguladores sem tentar replicar suas funções. Suas avaliações apoiam a elaboração de políticas informadas e o debate público sem determinar soluções específicas. Essa abordagem colaborativa permite o aprimoramento das estruturas e mecanismos de supervisão dos sistemas de IA, preservando a separação de poderes essencial para a governança democrática.

- **Empoderando o debate público e a compreensão do público**

O impacto categórico do MOS vem de sua capacidade de esclarecer questões complexas para a compreensão do público. Por meio de reuniões públicas regulares, sessões de *feedback* da comunidade e informações publicas sobre suas operações, ele ajudará os cidadãos a entender e a se envolver com questões sobre o papel da IA na sociedade. Em vez de tomar decisões pelo público, ele capacita a participação pública informada em debates cruciais sobre como a IA molda nosso ambiente de informações e processos democráticos.

Essa abordagem cuidadosamente equilibrada garante que o MOS enriqueça a supervisão dos sistemas de IA sem ultrapassar os limites da censura ou da jurisdição do território. Mantendo limites claros, se permite fornecer perspectivas de maior clareza, e fortalecer, em vez de passar por cima, das instituições democráticas existentes.

Considerações finais

A iniciativa **AI Commons** representa uma visão que busca democratizar a governança da IA por meio de quatro pilares interconectados que atendem tanto às necessidades imediatas quanto às mudanças estruturais de longo prazo. Ao combinar a capacitação por meio da Rede de Laboratórios de Equidade de IA, a contribuição da comunidade por meio do Conselho Cidadão de Desenvolvimento (CDC), a experimentação de políticas no Laboratório de Inovação em Políticas de IA (APIL) e a supervisão transparente por meio do Sistema de Supervisão de Múltiplas Partes Interessadas (MOS), essa estrutura visa criar vários caminhos para uma participação pública significativa na formação do desenvolvimento da IA. É importante ressaltar que essa iniciativa tem como objetivo a projeção de perspectivas do Sul Global e aborda os desequilíbrios fundamentais de poder no atual ecossistema de IA, ao mesmo tempo em que desenvolve a expertise local e a capacidade de tomada de decisões. Essa abordagem abrangente reconhece que a governança eficaz da IA exige inovação técnica e transformação social - desde o tratamento das desigualdades raciais e geográficas até a garantia da soberania digital e a reconstrução de espaços sociais compartilhados. O sucesso da iniciativa dependerá de um compromisso contínuo com a governança inclusiva, processos transparentes e uma redistribuição genuína de poder para garantir que os sistemas de IA sirvam aos valores democráticos e aos direitos humanos, ao invés de prejudicá-los.