ANOTAÇÕES PARA GOVERNANÇA DE UMA DAO PARA AMPLIAR O DEBATE DA INTERAÇÃO HUMANO-ALGORITMO E DAS TECNOLOGIAS EMERGENTES COM SEUS IMPACTOS SOBRE A HUMANIDADE

Juliao Braga, UFABC, Santo André, SP, BR (juliao.braga@ufabc.edu.br)[1,2]
Francisco Regateiro, IST, ULisboa, PT (francisco.regateiro@ist.ulisboa.pt)[3,4]
Itana Stiubiener, UFABC, Santo André, SP, BR (itana.stiubiener@ufabc.edu.br)[5,6]
Juliana Cristina Braga, UFABC, Santo André, SP, BR (juliana.braga@ufabc.edu.br)[7,8]

Abstract. *Human-algorithm interaction is at the top of humanity's concerns in view of the repercussions of the recent release of ChatGPT3 and 4. In late March/2023 a manifesto was released and signed by numerous stakeholders and has taken the news in magazines, newspapers and television, accompanied by numerous papers and talks by experts. This work starts from the assumption that human-algorithm interactions crosses a context of global interests and the involvement of stakeholders from the most diverse areas of human knowledge, adding aspects of high complexity. In this way, the work proposes a solution for there to be an effective participation of stakeholders from the various areas involved and society, in an effective debate, generalizing to digital platforms, in general. The proposal involves the creation of an Autonomous Decentralized Organization (DAO), which will inhabit the blockchain environment of the Ethereum network.*

Resumo. *A interação humano-algoritmo está no topo das preocupações da humanidade tendo em vista as repercussões do lançamento recente do ChatGPT3 e 4. No final de março/2023 um manifesto foi lançado e assinado por inúmeras interessados e tem tomado as notícias de revistas, jornais e televisão, acompanhadas por inúmeros trabalhos e palestras de especialistas. Este trabalho parte do pressuposto de que as interações humano-algoritmo atravessam um contexto de interesses globais e o envolvimento de partes interessadas das mais diversas áreas do conhecimento humano, adicionando aspectos de complexidade elevada. Desta forma, o trabalho propõe uma solução para que haja uma participação efetiva de partes interessadas das diversas áreas envolvidas e a sociedade, em um debate efetivo, generalizando para plataformas digitais, em geral. A proposta envolve a criação de uma Organização Decentralizada Autônoma, que habitará o ambiente blockchain da rede Ethereum.*

KEYWORDS: ia, ai, dao, plataformas digitais, tecnologias emergentes

[1]LATTES: http://lattes.cnpq.br/7092085044582071
[2]ORCID: https://orcid.org/0000-0001-9542-3732
[3]https://fenix.tecnico.ulisboa.pt/homepage/ist13522
[4]ORCID: https://orcid.org/0000-0003-2229-4938
[5]LATTES: http://lattes.cnpq.br/4008970012663480
[6]ORCID: https://orcid.org/0000-0002-7149-4760
[7]LATTES: http://lattes.cnpq.br/7111526592323456
[8]ORCID: https://orcid.org/0000-0003-2385-0051

1. Introdução

Este trabalho é o primeiro resultado da proposta caracterizada em um artigo anteriormente escrito para propor o debate de ideias em torno da criação de um ambiente orientado a professores, pesquisadores, pensadores e outras diferentes partes interessadas, em governança de algoritmos de inteligência artificial (IA) e dados [1].

A principal motivação para a proposta de criação deste ambiente foi a característica complexa do problema. Regular ou estabelecer regras para desenvolvedores de algoritmos de IA é uma imensa tarefa que depende de um debate entre uma diversidade imensa de interessados pois envolve questões de natureza ética, social, humana, especifidades técnicas, política privada e pública, entre muitas outras áreas do conhecimento humano. Além do mais é um debate muito longo e que exige organização apropriada para garantir a persistência dos fatos produzidos por uma imensidão de partes interessadas, que pertencem a toda a humanidade.

Um problema com dificuldade parecida pode ser considerado: os protocolos que fazem a Internet funcionar apropriadamente e que garantem sua dinâmica funcionalidade, ao longo do tempo (passado, presente e futuro). Este exemplo emana do ecossistema em torno do *Internet Engineering Task Force* (IETF). Milhares de interessados se reúnem, presencialmente, três vezes ao ano e, no resto do ano, persistentemente, através de grupos de trabalho por e-mail. É uma organização imensa e eficaz que garante o funcionamento da Internet, como ela é hoje.

A Internet, com sua incontestável importância para a humanidade é a justificativa da existência de ecossistema próprio [2]. Governança de algoritmos de IA é um tema mais complexo do que a Internet. Há registros de que Governança de algoritmos de IA é, realmente, um tema mais complexo do que a Internet, porque envolve implicações, principalmente ofensivas, mas também mortais, diretamente ao ser humano. De um modo geral e independente da aplicação, estes sistemas são considerados uma caixa preta resultando em informações assimétricas entre os seus desenvolvedores e seus consumidores [3]. Um dos exemplos mais tristes e que evidenciam a consequência desta assimetria é o projeto do sistema MCAS[9] do Boeing 737 MAX, que levou a dois acidentes com 346 mortes em outubro de 2018 (Lion Air) e março de 2019 (Ethiopian Airlines). Quando o ângulo do sensor de ataque falhou, os algoritmos embutidos forçaram o avião a baixar o nariz, resistindo às repetidas tentativas dos pilotos, confusos, de virar o nariz para cima. Ben Shneiderman, em seu livro *Human-Centered AI*, que comenta os dois acidentes com o Boeing 737 MAX, considera que o futuro destes algoritmos de IA é centrado no ser humano, principalmente tornando-se super ferramentas, que amplificam as habilidades humanas, capacitando as pessoas de forma notável mas, ao mesmo tempo, garantindo o controle humano [4]. Ben nomeou estes algoritmos com a sigla **HCAI**, acrônimo do título de seu livro.

Existem inúmeras outras aplicações usando IA ou não. Por exemplo, aquelas que se hospedam na Internet e que se comportam de forma desproporcional. Descrição detalhada, como representação dos chamados vieses algorítmicos podem ser encontradas no livro de Safya Noble, Algoritmos da Opressão e em outros [5] [6] [7].

Informações assimétricas, vieses e outras questões pertinentes estão incomodando os desenvolvedores, pesquisadores e outras partes interessadas em descobrir o que está faltando [8]. Perspectivas associadas a ética [9] [10] [11] [12] [13], regulamentação [14] [15] [16] [17] [18], governança [19] [20] [21] [22] [23] [3] [24] e muitas outras [25] [26] [27] [28] [29] [30] [31] [32] [33] estão na pauta de todas as partes interessadas, em busca de alternativas apropriadas – por exemplo, estas questões estão amplamente debatidas em [4]. Há uma extensa

[9] Acrônimo de *Manoeuvring Characteristics Augmentation System*

literatura apresentada por tópicos no trabalho não publicado que deu origem às abordagens deste artigo [1].

Se tivéssemos uma motivação tão forte como tem a Internet (com sua imensa capilarização mundial), provavelmente o modelo do IETF, ambientado em uma ampla participação de partes interessadas, como se pode ver na Figura 1 em sua amplitude, seria uma solução que certamente iria resolver as questões envolvendo a questão da governança de algoritmos e dados.

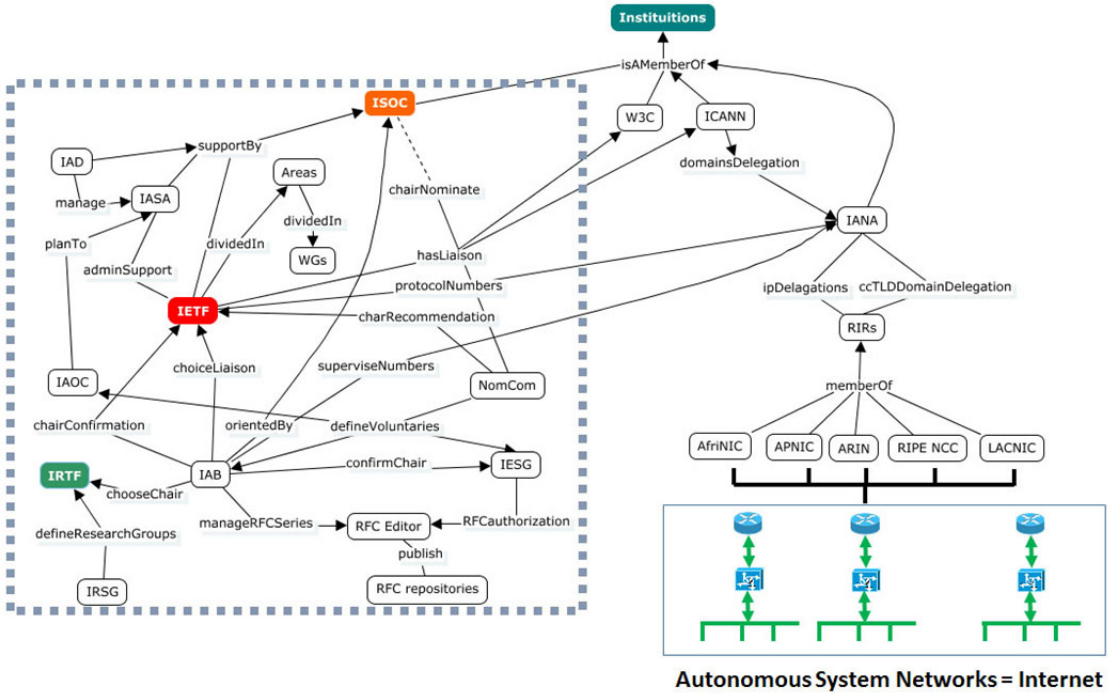

Figure 1. Ecosistema da governança da Internet. Fonte: [34]

Mas há sinais recentes nas atividades da IA que estão produzindo uma perspectiva de ampliar o seu conhecimento pela população mundial. Trata-se do sucesso imenso do ChatGPT, uma ferramenta que está popularizando a IA e ramificando seu protagonismo. Isto tem ocorrido com uma velocidade inimaginável apesar do fato de que o próprio ChatGPT propagar suas restrições e limitações sobre a compreensão da linguagem natural, na sua dependência de dados de treinamento e na possibilidade de vieses. Adicione ao ChatGPT e outras IAs generativas, o fato de possibilitarem a difusão de vieses. As IAs generativas são influenciadas pelos dados de treinamento levando a respostas tendenciosas ou discriminatórias. O perigo aumenta, com a possibilidade de aumento da base de treinamento durante a interação com o seu usuário, isto é, grupos com segundas intenções podem influenciar os dados de treinamento. O aperfeiçoamento das IAs generativas estão levando a expansão de suas possibilidades, para o uso de recursos multimodais, como é o caso do ChatGPT 4[10] ou o DALL-E 2[11]. De fato, a comunidade envolvida considera que após tais recentes lançamentos estamos em uma mudança significativa na IA. Bill Gates, por exemplo, reage através de um documento com sete capítulos, intitulado "The Age of AI has begun". No Capítulo 7 finaliza com "The Age of AI is filled with opportunities and responsibilities." [35]. Outras preocupações vem do neurocientista Miguel Nicolelis, em seu mais recente livro, "O verdadeiro criador de tudo: Como o cérebro humano

[10] https://openai.com/product/gpt-4
[11] https://openai.com/product/dall-e-2

esculpiu o universo como nós o conhecemos" expõe nos dois capítulos finais "...os graves riscos que a humanidade enfrentará nos próximos anos, em decorrência da nossa interação e da nossa dependência cada vez maiores em relação aos sistemas digitais, estabelecendo uma verdadeira simbiose que pode afetar profundamente o cérebro, por meio do fenômeno da plasticidade neural. Basicamente, a convivência quase contínua com computadores pode afetar a forma como o cérebro funciona e, no limite, nos transformar em meros zumbis digitais." [36]. Ademais, não se pode esquecer das histórias que volta e meio são contatas sobre as IAs generativas famosas e recentes [37].

Este trabalho propõe a criação de uma DAO denominada GHAIA DAO[12] como um mecanismo de governança original. Para desenvolver esta proposta, os autores criaram uma base de conhecimento sobre DAOs, que está disponível, incluindo suas atualizações, em um ambiente público do Open Science Framework (OSF) [38]. Adicionalmente, este trabalho propõe a criação da GHAIA DAO, cujo objetivo é suportar um ambiente de debates e registro de opiniões de partes interessadas nas questões que envolvem regulação de algoritmos de IA e dados bem como as questões envolvendo as interações humano-algoritmo, amplamente discutida na Cátedra Oscar Sala[13], do IEA-USP[14] cujo titular é o Prof. Dr. Virgílio Almeira (ORCID: 0000-0001-6452-0361[15]).

Além desta seção de Introdução, este trabalho discute sobre DAOs e sua variedade de governanças, na seção 2. Na seção 2.1, é apresentada a base de conhecimento construída no Protégé, um editor de ontologia gratuito e de código aberto, além de ser *famework* para a construção de sistemas inteligentes [39]. Ainda nesta seção é exibido as alternativas para uso desta base de conhecimento, em particular, orientando sobre o uso da linguagem ***SPARQL*** como ferramenta de pesquisas em ontologias [40, 41]. Na seção 5 é apresentada a proposta para a criação da GHAIA DAO, a qual inclui um mecanismo original de governança. Na seção 7 uma coleção de literatura complementar, classificada por áreas de interesse. Na seção 8 o trabalho aborda as conclusões desta etapa do trabalho e recomenda futuras atividades a serem seguidas. Na seção 9, os agradecimentos e na sequência, a relação bibliográfica.

2. DAOs

Uma *Decentralized Autonomous Organization* (DAO) é uma forma de organização baseada em *blockchain*, que geralmente é governada por seus membros, detentores de *tokens*. *Tokens*, uma espécie de criptomoedas (entre outros significados) podem ser adquiridos ou recebidos de alguma forma, por qualquer pessoa que, como proprietária destes *tokens* ganha o direito de votar em assuntos diretamente relacionados à governança da DAO. As regras de governança das DAOs são caracterizadas através de programas de computador conhecidos como ***contratos inteligentes*** (*smart contracts*), os quais são executados e validados dentro da *blockchain* da rede Ethereum[16], através de um recursos chamado /textitEthereum Virtual Machine (EVM). As características dos contratos inteligentes, como um banco de dados distribuído da *blockchain*, fazem com que as regras da organização sejam aplicadas pelo próprio código que define a DAO, tornando-a assim, ***autogovernada***.

Portanto, as DAOs são diferentes das empresas tradicionais porque são organizações autogovernadas, que funcionam de forma autônoma e descentralizada, sem a necessidade

[12]https://ghaia.pt
[13]https://bit.ly/cos-usp
[14]http://www.iea.usp.br/
[15]https://orcid.org/0000-0001-6452-0361
[16]https://ethereum.org/pt-br/

de intermediários. Enquanto que, as organizações tradicionais são sujeitos a direitos e responsabilidades definidos pelo sistema legal do ambiente no qual operam.

2.1. Conhecendo com detalhes sobre DAOs, implementados em uma ontologia

São muitos os tipos, funções e características das DAOs, entre aproximadamente 150 implementações na rede Ethereum, até abril/2023. Considerando tal diversidade, decidiu-se por criar uma base de conhecimento (KB[17]), através de uma ontologia, para auxiliar no entendimento e manter permanente as informações detalhadas sobre elas. Para criar esta base de conhecimento foi usado o software Protégé [39]. O Protégé, desenvolvido pela Universidade de Stanford[18] é um editor de ontologias gratuito e de código aberto e, complementarmente, um sistema de gerenciamento de conhecimento.

A ontologia criada para efeitos deste trabalho está disponível no ambiente público do projeto no *Open Science Framework* (OSF) [38]. A Figura 2 mostra as classes e seus relacionamentos, utilizados para a construção da referida ontologia, armazenada em ***decom.ttl***, no formato *Turtle* produzido pelo Protégé após a criação.

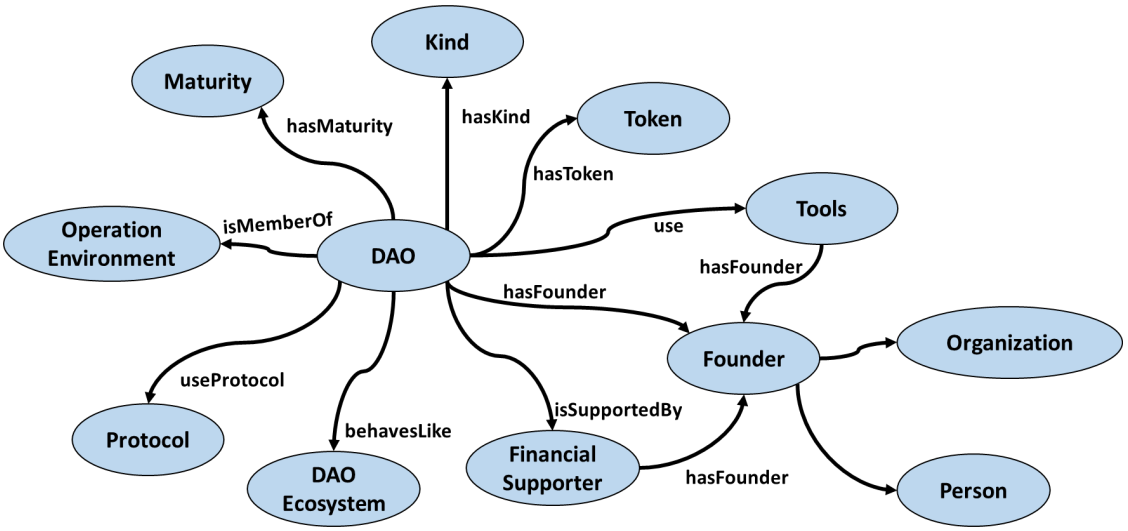

Figure 2. Termos, isto é, classes, subclasses e os respectivos relacionamentos que produziram mais de 4.000 itens na ontologia construída.

As Figuras 3 e 4 exibem, respectivamente, a classe auxiliar ***Significado de Acrônimos*** (*Acronym Meaning*) e as subclasses de ***Ferramentas*** (*Tools*).

Os detalhes referentes à construção da ontologia estão contidas em artigo próprio, em fase de preparação.

3. Como Pesquisar a Base de Conhecimento

Ao terminar a construção da **KB**, com um número substancial de axiomas, o principal interesse do ser humano se volta para a pesquisa desta **KB**. Umas das ferramentas para isto é a conhecida linguagem ***SPARQL*** *SPARQL Protocol and RDF Query Language* [42] [43] [44]. O Protégé oferece facilidades para usarmos a ***SPARQL*** e, também, o Apache Jena[19] [45]. Além destes dois exemplos, a DBpedia[20] e a Wikidata[21] disponibilizam interfaces públicas para a ***SPARQL***,

[17] Acrônimo do inglês: *knowlwdge Base*
[18] `https://protege.stanford.edu/`
[19] `https://jena.apache.org/`
[20] `http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VOSSPARQL`
[21] `https://w.wiki/rL`

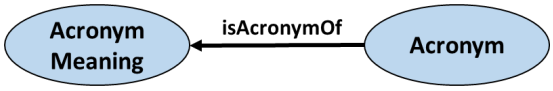

Figure 3. Acrônimos

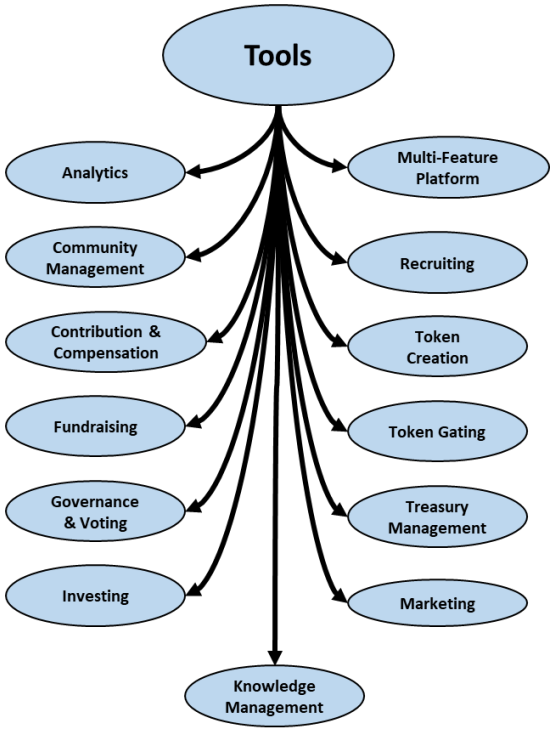

Figure 4. Subclasses de Tools

entre muitas outras. Em todos as ferramentas utilizadas, exceto o Protégé, usa-se como *entry point*, a URL `https://ghaia.pt/kb/decom.ttl`. No Protégé, a SPARQL age sobre a ontologia que está carregada.

O Protégé oferece outros recursos para pesquisar a ontologia produzida através dele. Por exemplo, o ***OntoGraph***, que oferece imagens como a Figura 5 e o recurso.

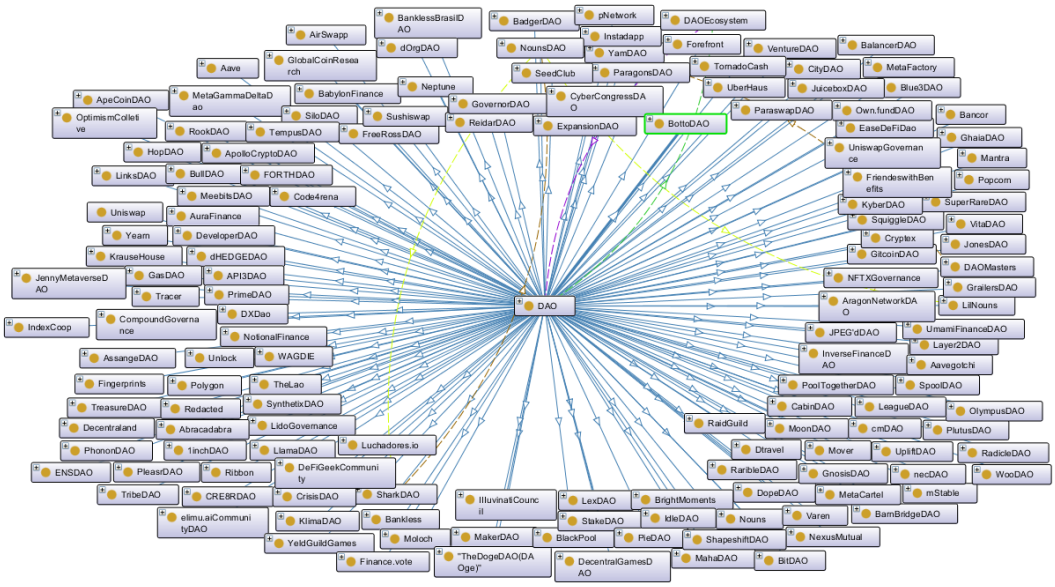

Figure 5. Visão gráfica das DAOs da base de conhecimento

Outro exemplo, o uso do recurso DL Query permite resultados imediatos a consultas usando conectivos lógicos, inclusive (Figura 6).

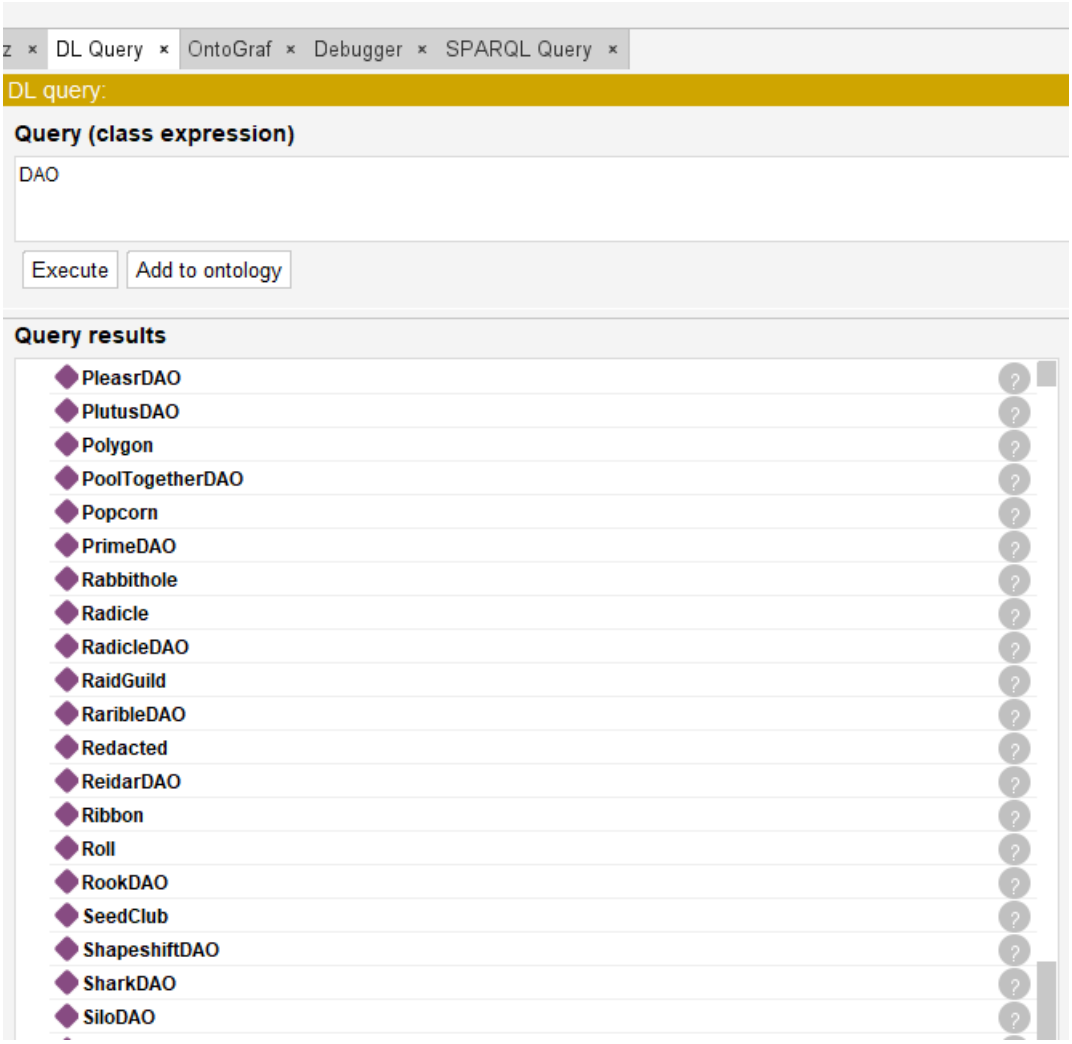

Figure 6. Exemplo de uma pesquisa e respectivo resultado parcial, usando o recurso DL Query, do Protégé.

4. Algoritmos de IA e seus vieses

A Cátegra Oscar Sala, sob a Coordenção do Prof. Dr. Virgílio Almeida se preocupou com "Algoritmos, inteligência artificial (IA), robôs e máquinas operadas por algoritmos que mediam cada vez mais nossas interações sociais, culturais, econômicas e políticas." em seu Projeto **Interações Humano-Algoritmo**.

Para a Cátedra[22] "existem três motivações básicas para o foco proposto para a Cátedra Oscar Sala em 2022/23:

1. Há diversos tipos de algoritmos em operação e com um papel cada vez maior na sociedade e nas atividades diárias dos cidadãos.
2. A complexidade dos algoritmos e dos sistemas que integram vários algoritmos vem aumentando rapidamente, novos modelos e gigantescas massas de dados tornam esses algoritmos e sistemas opacos, dificultando sobremaneira a compreensão do comportamento dos mesmos.

[22] `https://bit.ly/cosvirgilioalmeida`

3. A compreensão dos impactos sociais, políticos e econômicos, sejam eles positivos ou negativos, é um desafio de pesquisa."

Prof. Virgílio se antecipou a uma preocupação global com tais algoritmos opacos que concentram em sua maior parte naqueles de IA, se provou ser uma questão de tamanha prioridade para a humanidade culminando, muito recentemente, com um manifesto produzido pelo *Future of Life Institute* em relação às preocupações do ChatGPT (*Chat Generative Pre-trained Transformer*) [46], propondo uma moratória de seis meses no treinamento dele e de outros algoritmos parecidos, principalmente aqueles que usam grandes modelos de linguagem [47].

Livros, artigos científicos, artigos em jornais, e muitas outras formas de expressão dispuseram suas opiniões preocupações e recomendações para afrontar a invasão desregulada dos algoritmos com vieses de todos os tipos. Há ofensas graves e até mesmo ofensas criminosas. Mas, os textos antes do memorial descrito no parágrafo anterior foram intensos.

Além dos documentos acima, muitos outros, principalmente recentes, com suas referências abordam a mesma questão [48] [49] [50] [51] [52] [53] [54] [55] [56] [57] [58] [59] [60].

A proposta da Cátedra Oscar Sala, para as atividades do ano 2022/2033 mostrou sua imensa diversidade acadêmica e, na opinião dos autores, um assunto de complexidade bastante elevada merecendo uma atenção especial e debate contínuo com participação de grande número de partes interessadas, com uma forma de governança adequada, copiando a proposta implementada no IETF pela Internet Societ.

Portanto, com base nas preocupações da Internet Society (ISOC) sobre o ecossistema que governa a Internet, os autores propõe a criação de uma DAO, denominada GHAIA DAO e apresentada nas seções que seguem.

5. Constituição da GHAIA DAO

Uma das principais preocupação de uma DAO é a sua governança, que deve ser eficiente e democrática. As primeiras DAOs se propuseram a admitir que sua governança fosse feita pelos seus membros, que detinham os *tokens*. Um membro da DAO teria uma quantidade de votos equivalente ao número de *tokens* de sua propriedade. Com o passar do tempo verificou-se de que interessados em controlar as votações, induzindo a governança para da forma como quisessem, podeia fazer isto facilmente, comprando número suficiente de *tokens*. E isto começou a acontecer, e tais participantes eram chamados de ***baleias***.

Na tentativa de aperfeiçoar o processo de governança usou-se um esquema de votação no qual cada participante teria um voto. Os ***baleias*** não deram trégua e criaram participantes fantasmas ou usaram procurações, para aumentar o poder de suas participações. Um novo esquema foi adotado, o denominado esquema quadrático, garantindo que a votação seria decentralizada [61]. Nesta proposta, a votação continua por membro, mas cada membro recebe um número de votos equivalente ao dobro de seus *tokens*, que podem, adicionalmente ser usados na votação. Suponha que o membro ***A*** tenha 5 *tokens* e o membro ***B*** tenha 10 *tokens*. Então, ***A*** terá 10 *tokens* que poderá usar 5 como créditos para votar em uma proposta e 5 créditos para votar em outra proposta. Já o membro ***B*** terá 100 créditos podendo usar 50 créditos para uma proposta e 50 créditos em outra proposta. O membro que tem mais *tokens* pode gastar os seus *tokens* em uma proposta, mas se o outro membro tiver mais membros apoiando sua proposta, ele pode ganhar a votação. O voto quadrático permite que os usuários “paguem” por votos adicionais em uma determinada proposta para expressar com

mais força seu apoio a determinadas questões, resultando em resultados de votação alinhados com a maior disposição de participar (ou pagar...), em vez de apenas o resultado preferido pela maioria, independentemente da intensidade das preferências individuais. Esta questão sobre a democracia exercida em uma DAO leva aos dilemas fundamentais das sociedades com seus paradoxos e comportamento dos indivíduos que as constituem, em particular naqueles que vivem nas sociedades em que funcionam instituições democráticas [62].

Outras DAOs, adotando um dos critérios acima estabelecem um Conselho de Governança que irá cuidar da sua governança durante um período previamente acordado.

6. A GHAIA DAO

DeSci (*Decentralized Science*) é um movimento recente, que visa utilizar novas tecnologias, como *blockchain* ou ***Web3***[23], para abordar alguns pontos problemáticos da pesquisa científica, silos e gargalos. é uma alternativa aberta e global ao sistema científico moderno que enfrenta muitos desafios. Ele estende a ideia de ciência aberta, permitindo que os cientistas levantem fundos, compartilhem dados experimentais e obtenham ideias. Um dos exemplos mais interessantes é o de uma ***DeSci*** para adequar a revisão por pares [63] [64] [63] [65] [66] [67].

A GHAIA DAO é uma ***DeSci*** e para sua governança usará o ORCID ID, sobre o qual descreve-se na seção seguinte.

6.1. ORCID

ORCID significa *Open Researcher and Contributor ID* e, é uma organização global, sem fins lucrativos, sustentada por honorários de suas organizações membros. Formam uma comunidade construída e governada por um Conselho de Administração representativo de membros com ampla representação de partes interessadas. [68].

O ***ORCID ID*** é um identificador único[24], persistente e gratuito para que indivíduos usem enquanto se envolvem em atividades de pesquisa, bolsas de estudo e inovação. O ***ORCID*** oferece um conjunto de Interfaces de Programação de Aplicações (APIs).

No início de abril de 2023, as estatísticas[25] do ***ORCID*** indicavam algo em torno de 9 milhões, quatrocentos e dez pesquisadores inscritos, distribuídos por 56 países. O Brasil era o terceiro país com mais inscritos (361.900), após os Estados Unidos (794.493) e China (412.925).

6.2. A governança da GHAIA DAO

A GHAIA DAO, quando implementada será uma DeSci para disponibilizar um ambiente de debates multidisciplinar entre partes interessadas nas questões que envolvem a interação humano algoritmo e que seguirá, em parte, o modelo de debates do IETF e IRTF[26] (*Internet Research Task Force*. As seguintes regras irão modelar a GHAIA DAO:

(a) Ela terá dois tipos de *tokens*: ***OR*** e ***NOR*** ambos, inicialmente, com o valor de 1 (um) USD[27].
(b) Seus membros serão de dois tipos: aqueles que possuem o ***ORCID*** e aqueles sem ***ORCID***.

[23] Denominação genérica e não muito aceitável, dada ao ambiente *blockchain*
[24] `https://bit.ly/orcid-idbilder`
[25] `https://info.orcid.org/orcid-statistics/`
[26] `htps://irtf.org`
[27] 1 (um) dólar americano

(c) As partes interessadas com ***ORCID*** receberão 1 ***OR***, gratuitamente. O ***OR*** equivale a 1 ***NOR***, que na data de implementação da GAIA DAO será equivalente a 1 (um) USD. As partes interessadas sem ***ORCID*** precisam de adquirir ***NORs*** ao valor de mercado, para serem membros.
(d) Um ***OR*** equivale a um ***NOR***. ***ORs*** não são comercialisáveis, mas os ***NORs*** são comercializáveis livremente.
(e) O detentor de um ***OR*** terá direito de receber um ***NOR*** 21 (vinte e hum) meses depois de ter recebido o ***OR*** sem perder o direito de votar.
(f) O tesouro da GHAIA DAO deverá possuir um colateral equivalente ao valor do números de ***NORs*** distribuídos em ***OR***. Em outras palavras, não se pode distribuir um ***OR*** sem que haja a garantia equivalente no tesouro da GHAIA DAO.
(g) Os participantes detentores de ***OR*** votam e podem ser votados para um conselho de 21 (vinte e um) membros, o qual cuidará da governança da GHAIA DAO.
(h) Os participantes não detentores de ***OR***, isto é, sem ***ORCID*** não posuem direito a voto, mas podem ser votados.
(i) Os participantes com ***OR*** também votam para o Conselho de Controle, composto de sete (7) membros, cujo objetivo é garantir que não haja excessos por parte do Conselho de Governança. Vide Figura 7.
(j) Fora do *blockchain*, o Conselho de Governança, através do Suporte Técnico manterá listas de e-mail equivalente aos grupos de trabalho do IETF/IRTF (WGs[28]) e outras semelhanças, para tornar efetivo o debate em torno das **Interações Humano-Algoritmo**.
(k) O ambiente da Internet da GHAIA DAO deve hospedar um repositório de documentos equivalentes às RFCs[29] (*Request for Comments*) do IETF que serão desenvolvidos por seus membros.
(l) Outras regras relacionadas ao comportamento social e ético de ambos os participantes deverão ser definidas.
(m) o Suporte Técnico e Operacional é composto de pessoal técnico, administrativo e outros, remunerado adequadamente.

A Figura 7 exibe a governança proposta para a GHAIA DAO.

Esta figura se abstrai de detalhes de implementação na rede Ethereum e dos recursos da Internet fora da *blockchain*, necessários a atender os objetivos da GHAIA DAO.

7. Literatura Relacionada

Na Tabela 1 está referenciada a literatura usada para que se possa entender o mecanismo de governança de algoritmos e dados e, permitir que haja um comparativo, também, das propostas recomendadas. A maior parte destas referências foram originalmente coletadas na proposta elaborada por exigência da fase preliminar da Cátedra Oscar Sala e não publicado [1].

As referências estão classificadas em oito categorias e não se esgotam na relação apresentada nesta proposta:

1. **Internet:** Incluem as referências que abordam o tema de governança da Internet.
2. **Algoritmos:** São as referências que exibem algoritmos de IA em diversas áreas de aplicação.
3. **DAO:** Referências que abordam as DAOs e respectivas técnicas sobre as quais elas são construídas (blockchain e criptomoedas).

[28] Work Groups
[29] `https://rfc-editor.org`

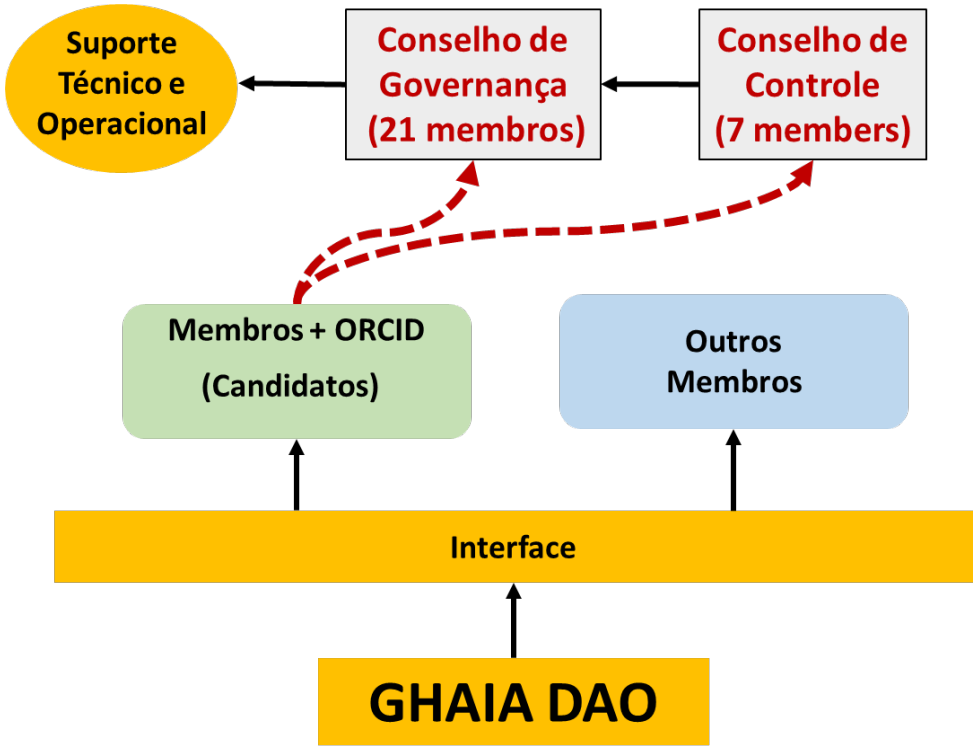

Figure 7. Estrutura de governança proposta da GHAIA DAO

Table 1. Estudos primários e secundários. por assunto

#	Referências	Classificação
1.	[69],[70],[71],[72],[73],[74],[75],[76],[77],[78], [79],[80] [81]	Internet
2.	[82],[83],[84],[85],[86],[87],[88],[89],[90],[91],[92],[93],[94],[95],[96],[97]	Algoritmos
3.	[98],[99],[100],[101],[102],[103],[104],[105],[106],[107],[108],[109],[110]	
4.	[111],[112],[113], [114],[115],[116], [117],[118],[119],[120],[121],[122],[123]	
5.	[124],[125],[126],[127],[128],[129],[130],[131],[132],[133],[134],[135],[136]	
6.	[137],[138],[139],[140],[141],[142][143],[144],[145],[146],[147],[148],[149]	
7.	[150],[151],[152],[153],[154],[155][156],[157],[158],[159],[160],[161],[162]	
8.	[163],[164],[165],[166],[167],[168],[169],[170],[171],[172],[173],[174],[175]	
9.	[176],[177],[178],[179],[180],[181],[182],[183],[184],[185],[186],[187]	
10.	[188],[189],[190],[191],[192],[193],[194],[195],[196],[197],[198],[199]	
11.	[200],[201],[202],[203],[204],[205],[206]	
12.	[207],[208],[209],[210],[211],[212],[213],[214],[215],[216],[217],[218]	DAO
13.	[219],[220],[221],[222],[223],[224],[225],[226],[227],[228],[229],[230],[231]	Economia
14.	[232],[233],[234],[235],[236],[237],[238],[239],[240],[241],[242],[243],[244]	
15.	[245]	
16.	[246],[247],[248],[249],[250],[251],[252],[253],[254],[255],[256],[257],[258]	Outros
17.	[259],[260],[261],[262],[263],[264],[265],[266],[267],[268],[269],[270],[271]	
18.	[272],[273],[274],[275],[276],[277],[278],[279],[280],[281],[282],[283],[284]	
19.	[285],[286],[287],[288],[289],[290],[291],[292],[293],[294],[295],[296],[297]	
20.	[298],[299],[300],[301],[302],[303], [304], [305]	
21.	[306],[15],[307],[308],[309],[310],[311],[312],[313],[314],[315],[316],[317]	Governança
22.	[318],[319],[320],[321],[322],[323],[324],[325],[326],[327],[328],[329],[330]	
23.	[331],[332],[333],[334],[335],[336],[337],[338],[339],[340],[341],[342],[343]	
24.	[344],[345],[346],[347],[348],[349],[6],[350],[351],[352],[353],[354],[7]	
25.	[355],[356],[357],[358],[359],[360],[361],[362],[58],[363],[364],[365],[366]	
26.	[367],[368],[369],[370],[371],[372],[373],[374],[375],[376],[377],[378],[379]	
27.	[380],[381],[382],[383],[384],[385],[386],[387],[388],[389],[390],[391],[392]	
28.	[393],[394],[395],[396],[397],[398]	
29.	[399],[400],[401],[402]	RLiteratura
30.	[403],[404],[405],[406],[407],[408],[409],[410],[25],[411],[412],[413],[414]	Social
31.	[415],[416],[417],[418],[419],[420],[421],[422],[423],[424], [13]	

4. **Economia:** Referências que abordam questões relacionadas com a economias dos algoritmos e seus ambientes.

5. **Outros:** Um conjunto de referências que descrevem o envolvimento de algoritmos de IA aos assuntos: Bots, Discriminação, Engenharia de Software, IA, Jogos, Robótica e Segurança.
6. **RLiteratura:** Trabalhos de revisão de literatura, incluindo revisões sistemáticas.
7. **Social:** Textos que referenciam os aspectos social, ético e filosófico dos algoritmos.

8. Conclusões e trabalhos futuros

Há muito trabalho a ser feito para a constituição da GHAIA DAO. A indisponibilidade atual de recursos impediu a implementação da GHAIA DAO, mas considerou-se um empecilho momentâneo. A KB de DAOs mostrou ser uma solução apropriada para aprender sobre DAOs, simplificando sua apresentação.

Na sequência considerou-se as seguintes tarefas como mandatórias, para o futuro:

(i) O ***ORCID*** deve ser informado das intenções em usar o ***ORCID ID*** para identificar os futuros membros votantes da GHAIA DAO.

(ii) O modelo econômico da estrutura de funcionamento da GHAIA DAO deve ser construído formalmente antes de sua constituição. Este modelo, entre muitos outros resultados, deve estimar o colateral seguro para seu início. O parâmetro base para esta formulação, no início de abril de 2023 são os 9.410.000 pesquisadores registrados no ORCID, além dos custos envolvidos na manutenção do Suporte Técnico e Operacional. Espera-se que interessados no projeto possam desenvolver documentos nesta direção.

(iii) Após a definição do modelo econômico e estabelecidas todas as regras de funcionamento da DAO será desenvolvido um ou mais *smart contracts* para garantir a ***autogovernança*** da GHAIA DAO.

(iv) As DAOs ainda não são regulamentadas em muitos países, o que pode gerar incerteza jurídica. Partes interessadas, com especialidade em Direito, em particular Direito Internacional deveriam estudar esta questão.

(v) A ontologia criada através do Protégé e armazenada em ***decom.ttl*** deve ser terminada e comparada com a ontologia completa (***decom.owl***).

(vi) Um texto detalhando o uso da ***SPARQL*** sobre as duas bases deve ser desenvolvido em forma de tutorial, para difundir o trabalho desenvolvido e útil para a comunidade interessada.

(vii) Um trabalho complementar, apresentando os gráficos produzidos extensivamente pelo Protégé está disponível no OSF do projeto[30].

(viii) É apropriado, para o sucesso do projeto a presença, sem limites, de partes interessadas das mais variadas e imensas áreas do conhecimento, incluindo a sociedade em geral.

(ix) È de se esperar, que a DAO hospede interessados em ampliar as interações humano-algoritmos para o contexto de todas as plataformas digitais.

9. Agradecimentos

Os autores agradecem ao Professor Dr. Virgílio de Almeida, catedrático da Cátedra Oscar Sala, do Instituto de Estudos Avançados da USP e à Professora Dra. Paola Cantarini, coordenadora do Grupo 3 da Cátedra que viabilizaram o presente trabalho.

References

[1] Juliao Braga, Francisco Regateiro, Itana Stiubiener, and Juliana C Braga. Uma proposta para impulsionar as pesquisas em governança de algoritmos de ia e dados. English version: https://osf.io/sr7kt, Setembro 2022.

[30] `https://bit.ly/daoKBinGraphics`

[2] Wolfgang Kleinwächter and Virgilio AF Almeida. The internet governance ecosystem and the rainforest. *IEEE Internet Computing*, 19(2):64–67, 2015.
[3] Urs Gasser and Virgilio A.F. Almeida. A Layered Model for AI Governance. *IEEE Internet Computing*, 21(6):58–62, 2017.
[4] Ben Shneiderman. *Human-Centered AI*. Oxford University Press, 2022.
[5] Safya. Noble and Felipe Damorim. *Algoritmos da Opressão: Como os mecanismos de busca reforçam o racismo*. Editora Rua do Sabão, Rio de Janeiro, 1 edition, 2022.
[6] Cathy O'Neil, editor. *Algoritmos de Destruição em Massa*. Editora Rua do Sabão, Santo Andrá, SP, 2020.
[7] Magaly Prado. *Fake news e inteligência artificial: O poder dos algoritmos na guerra da desinformação*, volume 1. Almedina Brasil, 2022.
[8] M Mitchell Waldrop. What are the limits of deep learning? *Proceedings of the National Academy of Sciences*, 116(4):1074–1077, 2019.
[9] Katina Michael, Diana Bowman, Meg Leta Jones, and Ramona Pringle. Robots and socio-ethical implications [guest editorial]. *IEEE Technology and Society Magazine*, 37(1):19–21, 2018.
[10] Andrea Censi, Konstantin Slutsky, Tichakorn Wongpiromsarn, Dmitry Yershov, Scott Pendleton, James Fu, and Emilio Frazzoli. Liability, ethics, and culture-aware behavior specification using rulebooks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8536–8542. IEEE, 2019.
[11] Michael Kearns and Aaron Roth. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press, 2019.
[12] Flávio S. Corrêa Silva and Nina S. T. Hirata. Inteligência Ética. *Computação Brasil*, 7:15–18, 2022.
[13] Paola Cantarini. Por uma Ética da Inteligência Artificial com base na Poietica. *Revista Juridica*, 4(71):892–912, 2022.
[14] Christianr Mölle and Arnaud Amouroux, editors. *Governing the Internet: Freedom and Regulation in the OSCE Region*. Organization for Security and Co-operation in Europe (OSCE), 2007.
[15] Eduardo Bismarck. PL 21/2020, 2020.
[16] Europe Commission. Coordinated Plan on Artificial Intelligence 2021 Review. Technical report, European Commission, Brussels, 2021.
[17] Margrethe Vestager and Thierry Breton. Uma Europa Preparada para a Era Digital : Comissão propõe novas regras e ações para promover a excelência e a confiança na inteligência artificial. Technical report, Comissão Euroeia, 2021.
[18] Sociedade Brasileira de Computação. Ética e regulação na inteligência artificial. Technical report, Sociedade Brasileira de Computação, 7 2022.
[19] Solon Barocas, Sophie Hood, and Malte Ziewitz. Governing Algorithms: A Provocation Piece. *SSRN Electronic Journal*, pages 1–12, 2013.
[20] Florian Saurwein, Natascha Just, and Michael Latzer. Governance of algorithms: Options and limitations. *Info*, 17(6):35–49, 2015.
[21] Danilo Doneda and Virgilio A.F. Almeida. What Is Algorithm Governance? *IEEE Internet Computing*, 20(4):60–63, 2016.
[22] Lucas D. Introna. Algorithms, Governance, and Governmentality: On Governing Academic Writing. *Science Technology and Human Values*, 41(1):17–49, 2016.
[23] Martin Ebers and Marta Cantero Gamito, editors. *Algorithmic Governance and Governance of Algorithms: Legal and Ethical Challenges*. Springer, 2021.
[24] Fernanda Bruno. Rastros digitais sob a perspectiva da teoria ator-rede. *Revista Famecos*, 19(3):681–704, 2012.

[25] Adam D.I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 111, pages 8788–8790, 2014.

[26] David Lazer. The rise of the social algorithm. *Science*, 348(6239):1090–1091, 2015.

[27] Matthew O Jackson. *The human network: How your social position determines your power, beliefs, and behaviors*. Pantheon Books, 2019.

[28] Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

[29] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. The social dilemma of autonomous vehicles. *Science*, 352(6293):1573–1576, 2016.

[30] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[31] Mengyi Wei and Zhixuan Zhou. Ai ethics issues in real world: Evidence from ai incident database. *arXiv preprint arXiv:2206.07635*, 2022.

[32] European Data Protection Supervisor (EDPS). Towards a new digital ethics. Technical Report September, European Organization, 2015.

[33] Anne Magaly de Paula Canuto. Ética no Uso de Dados Biométricos: Histeria ou uma Preocupação Coerente? *Computação Brasil*, 7:36–39, 2022.

[34] Juliao Braga, Jeferson Campos Nobre, Lisandro Zambenedetti Granville, and Marcelo Santos. Como Protocolos Inovadores são Criados e Adotados em Escala Mundial: Uma visão sobre o Internet Engineering Task Force (IETF) e a Infraestrutura da Internet. In Taisy Silva Weber and Claudia Aparecida Martins, editors, *Jornadas de Atualiza cão em Informática 2020*, page 45. Sociedade Brasileira de Computa cão, Cuiabá, MT Brazil, 2020. Available in: https://doi.org/10.5753/sbc.5728.3.2.

[35] Bill Gates. The Age of AI has begun, 2023. Last accessed 23 march 2023.

[36] Miguel Nicolelis. *O Verdadeiro Criador de Tudo: Como o Cérebro Humano Moldou o Universo Tal Como o Conhecemos*. ELSINORE, Brasil, 2023.

[37] Bruno Garattoni. GPT-4 tenta assumir o controle de outro computador – e digitar "como escapar" no Google. Super Interessante, Publicado em 23 mar 2023, 15h08, 2023. Last accessed 24 march 2023.

[38] Juliao Braga, Francisco Regateiro, and Itana Stiubiener. Human-algorithm: Governance, Sep 2022. Acessed in 03/09/2022.

[39] Mark A Musen. The protégé project: a look back and a look forward. *AI matters*, 1(4):4–12, 2015.

[40] Steve Harris and Andy Seaborne. SPARQL 1.1 Query Language. W3c working draft, W3C, march 2023. https://www.w3.org/TR/sparql11-query/.

[41] W3C. SPARQL, March 2023. [Online; accessed 24-March-2023].

[42] Andy Seaborne and Steven Harris. SPARQL 1.1 query language. W3C recommendation, W3C, March 2013. https://www.w3.org/TR/2013/REC-sparql11-query-20130321/.

[43] Bob DuCharme. *Learning SPARQL: querying and updating with SPARQL 1.1*. " O'Reilly Media, Inc.", 2013.

[44] Jedrzej Potoniec. Learning sparql queries from expected results. *Computing and Informatics*, 38(3):679–700, 2019.

[45] Apache Jena. Tutorial sparql. `https://jena.apache.org/tutorials/sparql_pt.html`, 2023.

[46] OpenAI. Gpt-4 technical report. *arXiv*, 2023.

[47] Future of Life Institute. Pause giant ai experiments: An open letter, 2023.
[48] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.
[49] Edd Gent. A cryptocurrency for the masses or a universal id?: Worldcoin aims to scan all the world's eyeballs. *IEEE Spectrum*, 60(1):42–57, 2023.
[50] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2016.
[51] Benjamin S Bucknall and Shiri Dori-Hacohen. Current and near-term ai as a potential existential risk factor. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 119–129, July 2022.
[52] Joseph Carlsmith. Is power-seeking ai an existential risk? *arXiv preprint arXiv:2206.13353*, 2022.
[53] Brian Christian. *The Alignment Problem: Machine Learning and Human Values*. Norton & Company, 2020.
[54] Michael Cohen, Marcus Hutter, and M Osborne. Advanced artificial agents intervene in the provision of reward. *AI Magazine*, 43(3):282–293, 2022.
[55] T Eloundou et al. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*, 2023.
[56] Dan Hendrycks and Mantas Mazeika. X-risk analysis for ai research. *arXiv preprint arXiv:2206.05862*, 2022.
[57] Richard Ngo. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*, 2022.
[58] Stuart Russell. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.
[59] Max Tegmark. *Life 3.0: Being Human in the Age of Artificial Intelligence*. Knopf, 2017.
[60] Lukas Weidinger et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.
[61] Steven P Lalley and E Glen Weyl. Quadratic voting: How mechanism design can radicalize democracy. In *AEA Papers and Proceedings*, volume 108, pages 33–37, 2018.
[62] Paulo Trigo Pereira. *Prisioneiro, o amante e as sereias: instituições económicas, políticas e democracia*. Almedina, 2008.
[63] Bianca Trovò and Nazzareno Massari. Ants-review: A privacy-oriented protocol for incentivized open peer reviews on ethereum. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12480 LNCS, 2021.
[64] Sarah Hamburg. Call to join the decentralized science movement, 2021.
[65] Richard Walker and Pascal Rocha da Silva. Emerging trends in peer review - a survey. *Frontiers in Neuroscience*, 9, 2015.
[66] Richard Smith. Peer review: A flawed process at the heart of science and journals, 2006.
[67] Robert E. Gropp, Scott Glisson, Stephen Gallo, and Lisa Thompson. Peer review: A system under stress. *BioScience*, 67, 2017.
[68] ORCID. Open Researcher and Contributor ID, 2023. `https://info.orcid.org/what-is-orcid/`.
[69] Janet Abbate. *Inventing the Internet*. MIT Press, 1999.
[70] Lucas Andrade, Juliao Braga, Stefany Pereira, Rafael Roque, and Marcelo Santos. In-Person and Remote Participation Review at IETF. In *Proceeding of CSBC 2018 - V Workshop pre IETF*, page 11, Natal, RN Brazil,

July 2018. To be published. Available at: http://braga.net.br/papers/In-Person%20and%20Remote%20Participation%20Review%20at%20IETF.pdf.
[71] Lee A. Bygrave and Jon Bing. *Internet Governance: Infrastructure and Institutions*. Oxford University Press, New York, 2009.
[72] Diego Rafael Canabarro and Flavio Rech. A Governança da Internet: Definição, Desafios e Perspectivas. In *9o ENCONTRO DA ABCP*, page 17, 2014.
[73] Diego Rafael Canabarro. *Governança global da internet: tecnologia, poder e desenvolvimento*. Doutoral thesis, Federal University of Rio Grande do Sul, 2014.
[74] Alexandre Arns Gonzales. *Quem Governa a Governança da Internet? Uma análise do papel da Internet sobre os rumos do sitema-mundo*. Dissertação de mestrado, Universidade Federal do Rio Grande do Sul, 2016.
[75] Fabrício Pasquot Bertini Polido and Lucas Costa Dos Anjos, editors. *Marco Civil E Governança Da Internet: Diálogos Entre O Doméstico E O Global*. Faculdade de Direito da UFMG, 2016.
[76] Fabrício Bertini Pasquot Polido, Lucas Cosda dos Anjos, and Luíza Couto Chaves Brandão, editors. *Tecnologias e Conectividade: Direito e Políticas na Governança das Redes*. IRIS, 2017.
[77] Wolfgang Kleinwächter. The History of Internet Governance. In *Governing the Internet: Freedom and Regulation in the OSCE Region*, pages 41—-65. OSCE Region, Vienna, 2007.
[78] Christian Moller. Governing the Domain Name System: An Introduction to Internet Infrastructure. In *Governing the Internet: Freedom and Regulation in the OSCE Region*, pages 29–39. OSCE, Vienna, 2007.
[79] Alexandre Pacheco da Silva, Ana Paula Camelo, Diego R. Canabarro, and Flavio Rech Wagner, editors. *Estrutura e funcionamento da internet : aspectos técnicos, políticos e regulatórios*. FGV, 2021.
[80] Elizabeth Machado Veloso. Legislação sobre Internet no Brasil. Technical report, Camara dos Deputados, 2009.
[81] Janet Abbate*. L'histoire de l'Internet au prisme des STS. *Le temps des médias*, 18(1):170–180, 2012.
[82] Martin Adam, Michael Wessel, and Alexander Benlian. Ai-based chatbots in customer service and their effects on user compliance. *Electronic Markets*, 31(2):427–445, 2021.
[83] Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. Social fmri: Investigating and shaping social mechanisms in the real world. *Pervasive and mobile computing*, 7(6):643–659, 2011.
[84] Saar Alon Barkat and Madalina Busuioc. Human-ai interactions in public sector decision-making: Automation bias and selective adherence to algorithmic advice. *Accepted Manuscript*, 2022.
[85] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13, 2019.
[86] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 291–300. IEEE, 2019.
[87] Tawfiq Ammari, Jofish Kaye, Janice Y Tsai, and Frank Bentley. Music, search, and iot: How people (really) use voice assistants. *ACM Trans. Comput. Hum. Interact.*, 26(3):17–1, 2019.

[88] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[89] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias, 2016.

[90] George A. Akerlof and Rachel E. Kranton. *Identity Economics: How Our Identities Shape Our Work, Wages, and Well-Being*. Princeton University Press, Princeton, 1 edition, 2010.

[91] Kenneth C Arnold, Krysta Chauncey, and Krzysztof Z Gajos. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 128–138, 2020.

[92] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

[93] Lisanne Bainbridge. Ironies of automation. In *Analysis, design and evaluation of man–machine systems*, pages 129–135. Elsevier, 1983.

[94] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*, pages 1–14, 2018.

[95] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[96] Josh Bongard, Victor Zykov, and Hod Lipson. Resilient machines through continuous self-modeling. *Science*, 314(5802):1118–1121, 2006.

[97] N Bolstrom. *Superintelligence. Paths, dangers, strategies*. Oxford University Press, United Kingdom, 2014.

[98] Jeffrey M Bradshaw, Robert R Hoffman, David D Woods, and Matthew Johnson. The seven deadly myths of "autonomous systems". *IEEE Intelligent Systems*, 28(3):54–61, 2013.

[99] Elizabeth Broadbent. Interactions with robots: The truths we reveal about ourselves. *Annual review of psychology*, 68(1):627–652, 2017.

[100] Connor Brooks and Daniel Szafir. Visualization of intended assistance for acceptance of shared control. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11425–11430. IEEE, 2020.

[101] Connor Brooks and Daniel Szafir. Balanced information gathering and goal-oriented actions in shared autonomy. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 85–94. IEEE, 2019.

[102] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, et al. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. *arXiv preprint*, 2020.

[103] Erik Brynjolfsson and Tom Mitchell. What can machine learning do? Workforce implications. *Science*, 358(6370):1530–1534, 2017.

[104] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[105] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[106] Rafael A Calvo, Dorian Peters, Karina Vold, and Richard M Ryan. Supporting human autonomy in ai systems: A framework for ethical enquiry. In *Ethics of Digital Well-Being*, pages 31–54. Springer, 2020.

[107] Neil A. H. Campbell. The Evolution of Flight Data Analysis. In *Proceedings of Australian Society of Air Safety Investigators*, pages 1–22, 2003.

[108] Juliano Cappi. *Internet, Big Data e discurso de ódio: reflexões sobre as dinâmicas de interação no Twitter e os novos ambientes de debate político*. Doctoral thesis, Pontifícia Universidade Católica de São Paulo, 2017.

[109] Felix Carros, Johanna Meurer, Diana Löffler, David Unbehaun, Sarah Matthies, Inga Koch, Rainer Wieching, Dave Randall, Marc Hassenzahl, and Volker Wulf. Exploring human-robot interaction with the elderly: results from a ten-week case study in a care home. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[110] Radhika Chemuturi, Farshid Amirabdollahian, and Kerstin Dautenhahn. Adaptive training algorithm for robot-assisted upper-arm rehabilitation, applicable to individualised and therapeutic human-robot interaction. *Journal of NeuroEngineering and Rehabilitation*, 10(1), 2013.

[111] Bo Chen, Chunsheng Hua, Bo Dai, Yuqing He, and Jianda Han. Online control programming algorithm for human–robot interaction system with a novel real-time human gesture recognition method. *International Journal of Advanced Robotic Systems*, 16(4):1–18, 2019.

[112] Lu Cheng, Kush R Varshney, and Huan Liu. Socially responsible ai algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research*, 71:1137–1181, 2021.

[113] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148. PMLR, 2018.

[114] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pages 4300–4308, 2017.

[115] Jacob W Crandall, Mayada Oudah, Fatimah Ishowo-Oloko, Sherief Abdallah, Jean-François Bonnefon, Manuel Cebrian, Azim Shariff, Michael A Goodrich, Iyad Rahwan, et al. Cooperating with machines. *Nature communications*, 9(1):1–12, 2018.
textbfThis study examines algorithmic cooperation with humans and provides an example of methods that can be used to study the behaviour of human–machine hybrid systems.

[116] Antoine Cully, Jeff Clune, Danesh Tarapore, and Jean-Baptiste Mouret. Robots that can adapt like animals. *Nature*, 521(7553):503–507, 2015.

[117] Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–108, 2004.

[118] Rajarshi Das, James E Hanson, Jeffrey O Kephart, and Gerald Tesauro. Agent-human interactions in the continuous double auction. In *International Joint Conference on Artificial Intelligence*, volume 17, pages 1169–1178. Lawrence Erlbaum Associates Ltd, 2001.

[119] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

[120] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015.

[121] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):1–5, 2018.

[122] Douglas C Engelbart and William K English. A research center for augmenting human intellect. In *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*, pages 395–410, 1968.

[123] Douglas C Engelbart. Augmenting human intellect: A conceptual framework. *Menlo Park, CA*, page 21, 1962. Reprinted in Packer, R. amd Kprdam. L;. eds; (2–1). Multimedia: From Wagner to Virtual Reality. New York: W. W. Norton, 64-90.

[124] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Proceedings of Machine Learning Research*, pages 1–12, 2018.

[125] Ziv Epstein, Blakeley H Payne, Judy Hanwen Shen, Abhimanyu Dubey, Bjarke Felbo, Matthew Groh, Nick Obradovich, Manuel Cebrian, and Iyad Rahwan. Closing the ai knowledge gap. *arXiv preprint arXiv:1803.07233*, 2018.

[126] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2015.

[127] Sophie Freeman, Martin Gibbs, and Bjørn Nansen. 'don't mess with my algorithm': Exploring the relationship between listeners and automated curation and recommendation on music streaming services. *First Monday*, 2022.

[128] Tarleton Gillespie. The Relevance of Algorithms. In *Media technologies: Essays on communication, materiality, and society*, pages 167–194. The MIT Press, 2014.

[129] Enric Galceran, Alexander G Cunningham, Ryan M Eustice, and Edwin Olson. Multipolicy decision-making for autonomous driving via changepoint-based behavior prediction: Theory and experiment. *Autonomous Robots*, 41:1367–1382, 2017.

[130] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64:86–92, 2021.

[131] Alessandro Giusti, Jérôme Guzzi, Dan C Cireşan, Fang-Lin He, Juan P Rodríguez, Flavio Fontana, Matthias Faessler, Christian Forster, Jürgen Schmidhuber, Gianni Di Caro, et al. A machine learning approach to visual perception of forest trails for mobile robots. *IEEE Robotics and Automation Letters*, 1:661–667, 2015.

[132] Vern L Glaser. *Enchanted algorithms The Quantification of Organizational Decision-Making*. University of Southern California, 2014.

[133] Kurt Gray and Daniel M Wegner. Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, 125(1):125–130, 2012.

[134] Victoria Groom and Clifford Nass. Can robots be teammates?: Benchmarks in human–robot teams. *Interaction studies*, 8(3):483–500, 2007.

[135] Jaron Harambam, Dimitrios Bountouridis, Mykola Makhortykh, and Joris Van Hoboken. Designing for the better by taking users into account: A qualitative evaluation of user control mechanisms in (news) recommender systems. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 69–77, 2019.

[136] Jeffrey Heer. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences*, 116:1844–1850, 2019.

[137] Bram Hendriks, Bernt Meerbeek, Stella Boess, Steffen Pauws, and Marieke Sonneveld. Robot vacuum cleaner personality and behavior. *International Journal of Social Robotics*, 3:187–195, 2011.

[138] Martin Hilbert, Saifuddin Ahmed, Jaeho Cho, Billy Liu, and Jonathan Luu. Communicating with Algorithms: A Transfer Entropy Analysis of Emotions-based Escapes from Online Echo Chambers. *Communication Methods and Measures*, 12(4):260–275, 2018.

[139] Gunter J Hitsch, Ali Hortaçsu, and Dan Ariely. Matching and sorting in online dating. *American Economic Review*, 100(1):130–63, 2010.

[140] Shanee Honig, Alon Bartal, Yisrael Parmet, and Tal Oron-Gilad. Using online customer reviews to classify, predict, and learn about domestic robot failures. *arXiv preprint arXiv:2201.03287*, 2022.

[141] Shanee Honig and Tal Oron-Gilad. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in psychology*, 9:861, 2018.

[142] Xiyang Hu, Yan Huang, Beibei Li, and Tian Lu. Uncovering the Source of Evaluation Bias in Micro-Lending. In *ICIS 2021 Proceedings*, volume 1. Association for Computing Machinery, 2021.

[143] Yin-Fu Huang and Yi-Hao Li. Sentiment translation model for expressing positive sentimental statements. In *2019 International Conference on Machine Learning and Data Engineering (iCMLDE)*, pages 79–84. IEEE, 2019.

[144] Lillian Hung, Mario Gregorio, Jim Mann, Christine Wallsworth, Neil Horne, Annette Berndt, Cindy Liu, Evan Woldum, Andy Au-Yeung, and Habib Chaudhury. Exploring the perceptions of people with dementia about the social robot paro in a hospital setting. *Dementia*, 20:485–504, 2021.

[145] Nicholas R Jennings, Luc Moreau, David Nicholson, Sarvapali Ramchurn, Stephen Roberts, Tom Rodden, and Alex Rogers. Human-agent collectives. *Communications of the ACM*, 57:80–88, 2014.

[146] Sooyeon Jeong, Cynthia Breazeal, Deirdre Logan, and Peter Weinstock. Huggable: the impact of embodiment on promoting socio-emotional interactions for young pediatric inpatients. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

[147] Neil Johnson, Guannan Zhao, Eric Hunsader, Hong Qi, Nicholas Johnson, Jing Meng, and Brian Tivnan. Abrupt rise of new machine ecology beyond human response time. *Scientific reports*, 3(1):1–7, 2013.

[148] Ece Kamar, Severin Hacker, and Eric Horvitz. Combining human and machine intelligence in large-scale crowdsourcing. In *AAMAS*, volume 12, pages 467–474, 2012.

[149] William B Kannel and Daniel L McGee. Diabetes and cardiovascular disease: the framingham study. *Jama*, 241(19):2035–2038, 1979.

[150] Serge Kernbach, Ronald Thenius, Olga Kernbach, and Thomas Schmickl. Re-embodiment of honeybee aggregation behavior in an artificial micro-robotic system. *Adaptive Behavior*, 17:237–259, 2009.

[151] Peter H Kahn, Nathan G Freier, Takayuki Kanda, Hiroshi Ishiguro, Jolina H Ruckert, Rachel L Severson, and Shaun K Kane. Design patterns for sociality in human-robot interaction. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, pages 97–104, 2008.

[152] Hiroaki Kitano, Minoru Asada, Yasuo Kuniyoshi, Itsuki Noda, and Eiichi Osawa. Robocup: The robot world cup initiative. In *Proceedings of the first international conference on Autonomous agents*, pages 340–347, 1997.

[153] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 08 2017.

[154] Jacqueline M Kory Westlund, Sooyeon Jeong, Hae W Park, Samuel Ronfard, Aradhana Adhikari, Paul L Harris, David DeSteno, and Cynthia L Breazeal. Flat vs. expressive storytelling: Young children's learning and retention of a social robot's narrative. *Frontiers in human neuroscience*, 11:295, 2017.

[155] Johannes Kunkel, Claudia Schwenger, and Jürgen Ziegler. NewsViz: Depicting and Controlling Preference Profiles Using Interactive Treemaps in News Recommender Systems. *UMAP 2020 - Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 126–135, 2020.

[156] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-agent cooperation and the emergence of (natural) language. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pages 1–11, 2017.

[157] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[158] Michael L. Littman, Ifeoma Ajunwa, Guy Berger, Craig Boutilier, Morgan Currie, Finale Doshi-Velez, Gillian Hadfiel, Michael C. Horowitz, Charles Isbell, Hiroaki Kitano, Karen Levy, Terah Lyons, Melanie Mitchell, Julie Shah, Steven Sloman, Shannon Vallor, and Toby Walsh. Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report. *Stanford University, Stanford, CA*, pages 1–82, 2021.

[159] Irene Lopatovska, Katrina Rink, Ian Knight, Kieran Raines, Kevin Cosenza, Harriet Williams, Perachya Sorsche, David Hirsch, Qi Li, and Adrianna Martinez. Talk to me: Exploring user interactions with the amazon alexa. *Journal of Librarianship and Information Science*, 51(4):984–997, 2019.

[160] Gustavo López, Luis Quesada, and Luis A Guerrero. Alexa vs. siri vs. cortana vs. google assistant: a comparison of speech-based natural user interfaces. In *International conference on applied human factors and ergonomics*, pages 241–250. Springer, 2017.

[161] Tamara Lorenz, Astrid Weiss, and Sandra Hirche. Synchrony and reciprocity: Key mechanisms for social companion robots in therapy and care. *International Journal of Social Robotics*, 8(1):125–143, 2016.

[162] John Markoff. *Machines of loving grace: The quest for common ground between humans and robots*. HarperCollins Publishers, 2016.

[163] Sean McGregor. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, 17B:15458–15463, 2021.

[164] Michelle N Meyer. Two cheers for corporate experimentation: The a/b illusion and the virtues of data-driven innovation. *Colo. Tech. LJ*, 13:273, 2015.

[165] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[166] Alex Mitrevski, Santosh Thoduka, Argentina Ortega Sáinz, Maximilian Schöbel, Patrick Nagel, Paul G Plöger, and Erwin Prassler. Deploying robots in everyday environments: Towards dependable and practical robotic systems. *arXiv preprint arXiv:2206.12719*, 2022.

[167] Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 279–288, 2019.

[168] Alexandra S Mueller, Ian J Reagan, and Jessica B Cicchino. Addressing driver disengagement and proper system use: Human factors recommendations for level 2 driving automation design. *Journal of Cognitive Engineering and Decision Making*, 15(1):3–27, 2021.

[169] Simone Natale. To believe in Siri: A critical analysis of AI voice assistants. Technical Report March, University of Bremen, 2020.

[170] Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*, pages 677–686, 2014.

[171] Yuriy Nevmyvaka, Yi Feng, and Michael Kearns. Reinforcement learning for optimized trade execution. In *Proceedings of the 23rd international conference on Machine learning*, pages 673–680, 2006.

[172] Amit Kumar Pandey and Rodolphe Gelin. A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. *IEEE Robotics & Automation Magazine*, 25(3):40–48, 2018.

[173] Hae Won Park, Rinat Rosenberg-Kima, Maor Rosenberg, Goren Gordon, and Cynthia Breazeal. Growing growth mindset with a social robot peer. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, pages 137–145, 2017.

[174] Rik Peeters. The agency of algorithms: Understanding human-algorithm interaction in administrative decision-making. *Information Polity*, 25(4):507–522, 2020.

[175] Ola Pettersson. Execution monitoring in robotics: A survey. *Robotics and Autonomous Systems*, 53(2):73–88, 2005.

[176] Antonio Pérez, M Isabel García, Manuel Nieto, José L Pedraza, Santiago Rodríguez, and Juan Zamorano. Argos: An advanced in-vehicle data recorder on a massively sensorized vehicle for car driver behavior experimentation. *IEEE Transactions on Intelligent Transportation Systems*, 11(2):463–473, 2010.

[177] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, pages 97–101, 2016.

[178] Angel Rivas-Casado, Rafael Martinez-Tomás, and Antonio Fernández-Caballero. Multi-agent system for knowledge-based event recognition and composition. *Expert Systems*, 28(5):488–501, 2011.

[179] Florian Rosenberg and Schahram Dustdar. Design and implementation of a service-oriented business rules broker. In *Seventh IEEE International Conference on E-Commerce Technology Workshops*, pages 55–63. IEEE, 2005.

[180] Michael Rubenstein, Alejandro Cornejo, and Radhika Nagpal. Programmable self-assembly in a thousand-robot swarm. *Science*, 345(6198):795–799, 2014.

[181] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[182] Elliot Salisbury, Ece Kamar, and Meredith Ringel Morris. Toward scalable social alt text: Conversational crowdsourcing as a tool for refining vision-to-language technology for the blind. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*, 2017.

[183] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *science*, 311(5762):854–856, 2006.

[184] Filippo Santoni de Sio and Jeroen Van den Hoven. Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, page 15, 2018.

[185] Ben Shneiderman. The limits of speech recognition. *Communications of the ACM*, 43(9):63–65, 2000.

[186] Mehdi Shanbedi, Saeed Zeinali Heris, Ahmad Amiri, Sadegh Adyani, Mohsen Alizadeh, and Majid Baniadam. Optimization of the thermal efficiency of a two-phase closed thermosyphon using active learning on the human algorithm interaction. *Numerical Heat Transfer; Part A: Applications*, 66(8):947–962, 2014.

[187] Ben Sheehan, Hyun Seung Jin, and Udo Gottlieb. Customer service chatbots: Anthropomorphism and adoption. *Journal of Business Research*, 115:14–24, 2020.

[188] Donghee Shin. How do users interact with algorithm recommender systems? The interaction of users, algorithms, and performance. *Computers in Human Behavior*, 109(May 2019):106344, 2020.

[189] Hirokazu Shirado and Nicholas A Christakis. Locally noisy autonomous agents improve global human coordination in network experiments. *Nature*, 545(7654):370–374, 2017. **In this human-machine hybrid study, the authors show that simple algorithms injected into the human player can improve the results of human-to-human coordination.**

[190] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

[191] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[192] Wilko Schwarting, Javier Alonso-Mora, Liam Pauli, Sertac Karaman, and Daniela Rus. Parallel autonomy in automated vehicles: Safe motion generation with minimal intervention. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1928–1935. IEEE, 2017.

[193] Michel Taïx, David Flavigné, and Etienne Ferré. Human interaction with motion planning algorithm. *Journal of Intelligent and Robotic Systems: Theory and Applications*, 67(3-4):285–306, 2012.

[194] Andreas Theodorou, Robert H Wortham, and Joanna J Bryson. Designing and implementing transparency for real time inspection of autonomous robots. *Connection Science*, 29(3):230–241, 2017.

[195] Andrea L Thomaz and Cynthia Breazeal. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*, 172(6-7):716–737, 2008.

[196] Andreas Wagner. *Robustness and evolvability in living systems*. Princeton university press, 2013.

[197] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016.

[198] David Watson. The rhetoric and reality of anthropomorphism in artificial intelligence. *Minds and Machines*, 29(3):417–440, 2019.

[199] John Wenskovitch, Michelle Zhou, Christopher Collins, Remco Chang, Michelle Dowling, Alex Endert, and Kai Xu. Putting the "i" in interaction: Interactive interfaces personalized to individuals. *IEEE Computer Graphics and Applications*, 40(3):73–82, 2020.

[200] Jacqueline M Kory Westlund, Hae Won Park, Randi Williams, and Cynthia Breazeal. Measuring young children's long-term relationships with social robots. In *Proceedings of the 17th ACM conference on interaction design and children*, pages 207–218, 2018.

[201] Sangseok You and Lionel Robert. Emotional attachment, performance, and viability in teams collaborating with embodied physical action (epa) robots. *You, S. and Robert, LP (2018). Emotional Attachment, Performance, and Viability in Teams Collaborating with Embodied Physical Action (EPA) Robots, Journal of the Association for Information Systems*, 19(5):377–407, 2017.

[202] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[203] Xiaohua Zeng, Abraham O Fapojuwo, and Robert J Davies. Design and performance evaluation of voice activated wireless home devices. *IEEE Transactions on Consumer Electronics*, 52(3):983–989, 2006.

[204] Jie M Zhang, Mark Harman, Lei Ma, and Yang Liu. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering*, 2020.

[205] Eric Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 159–166, 1999.

[206] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2017, 2017.

[207] Stefano Angieri, Alberto García-Martínez, Bingyang Liu, Zhiwei Yan, Chuang Wang, and Marcelo Bagnulo. An experiment in distributed internet address management using blockchains. *arXiv preprint arXiv:1807.10528*, 2018.

[208] Juliao Braga, Joao Nuno Silva, Patricia Takako Endo, Jessica Ribas, and Nizam Omar. Blockchain to improve security, knowledge and collaboration inter-agent communication over restrict domains of the internet infrastructure, with human interaction. *Brazilian Journal of Development*, 5(7):9013–9029, july 2019. DOI:10.34117/bjdv5n7-103, ISSN 2525-8761.

[209] Igor M Coelho and Vitor N Coelho. Neocompiler eco: experimentaçao de consenso em blockchain e contratos inteligentes. In *Anais do VI Workshop do testbed FIBRE*, pages 57–67. SBC, 2021.

[210] Lara Bonemer Rocha Floriani. *Smart contracts nos contratos empresariais: um estudo sobre possibilidade e viabilidade econômica de sua utilização*. Editora Dialética, 2021.

[211] Alan E. Kazdin. *The token economy: A Review and Evaluation*. Plenum Press, 2012.

[212] Antony Lewis. *The Basics of Bitcoins and Blockchains: An Introduction to Cryptocurrencies and the Technology that Powers Them*. Group, Mango Publishing, Coral Gables, FL, 1 edition, 2018.

[213] Alex Murray, Dennie Kim, and Jordan Combs. The promise of a decentralized internet: What is web 3.0 and how can firms prepare? *Business Horizons*, 65:565–570, 2022.

[214] Ankita Saxena. Workforce Diversity: A Key to Improve Productivity. *Procedia Economics and Finance*, 11:76–85, 2014.

[215] Steven D. Travers. Distributed Autonomous Organization: A Blockchain Organizational Archetype. *Strategic Management Society 37th Annual Conference*, 2017.

[216] Shermin Voshmgir. *Economia dos Tokens: Como a Web3 está reinventando a internet e a relação entre os agentes econômicos*. Token Kitchen, 2 edition, 2020.

[217] Aries Wanlin Wang. *Crypto Economy: How Blockchain, Cryptocurrency and Token-Economy are Disrupting the Financial World*. Skyhorse Publishing, 2018.

[218] Guy R Vishnia and Gareth W Peters. Auditchain: A trading audit platform over blockchain. *Frontiers in Blockchain*, 3:9, 2020.

[219] W Brian Arthur. Complexity and the economy. In *Handbook of Research on Complexity*. Edward Elgar Publishing, 2009.

[220] Nicolas Bouleau. On excessive mathematization, symptoms, diagnosis and philosophical bases for real world knowledge. *Real World Economics*, 57:90–105, 2011.

[221] Kyle Croman, Christian Decker B, Ittay Eyal, Adem Efe Gencer, Ari Juels, Ahmed Kosba, Andrew Miller, Prateek Saxena, Dawn Song, and Roger Wattenhofer. On Scaling Decentralized Blockchains (A Position Paper). *Lecture Notes in Computer Science*, 9604:106–125, 2016.

[222] Eric Budish, Peter Cramton, and John Shim. The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quarterly Journal of Economics*, 130(4):1547–1621, 2015.

[223] John Cartlidge, Marco De Luca, Charlotte Szostek, and Dave Cliff. Too fast too furious: faster financial-market trading agents can give less efficient markets. In *ICAART-2012: 4th International Conference on Agents and Artificial Intelligence*, pages 126–135. SciTePress, 2012.

[224] Le Chen and Christo Wilson. Observing algorithmic marketplaces in-the-wild. *ACM SIGecom Exchanges*, 15(2):34–39, 2017.

[225] Le Chen, Alan Mislove, and Christo Wilson. An empirical analysis of algorithmic pricing on amazon marketplace. In *Proceedings of the 25th international conference on World Wide Web*, pages 1339–1349, 2016.

[226] J Doyne Farmer and Spyros Skouras. An ecological perspective on the future of computer trading. *Quantitative Finance*, 13:325–346, 2013.

[227] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Communications of the ACM*, 59:96–104, 2016.

[228] Michael Kearns, Alex Kulesza, and Yuriy Nevmyvaka. Empirical limitations on high-frequency trading profitability. *The Journal of Trading*, 5:50–62, 2010.

[229] Andrei A Kirilenko and Andrew W Lo. Moore's law versus murphy's law: Algorithmic trading and its discontents. *Journal of Economic Perspectives*, 27(2):51–72, 2013.

[230] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *Leibniz International Proceedings in Informatics, LIPIcs*, 67:1–23, 2017.

[231] Jon Kleinberg and Sigal Oren. Mechanisms for (mis) allocating scientific credit. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 529–538, 2011.

[232] Farshad Kooti, Mihajlo Grbovic, Luca Maria Aiello, Nemanja Djuric, Vladan Radosavljevic, and Kristina Lerman. Analyzing uber's ride-sharing economy. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 574–582, 2017.

[233] Jaron Lanier. *You are not a gadget*. Vintage, 2010.

[234] Michael Latzer, Katharina Hollnbuchner, Natascha Just, and Florian Saurwein. The economics of algorithmic selection on the internet. In *Handbook on the Economics of the Internet*, pages 395–425. University of Zurich, 2014.

[235] Albert J Menkveld. The economics of high-frequency trading. *Annual Review of Financial Economics*, 8:1–24, 2016.

[236] Miriam Naigembe. *Bank lending policy, credit scoring and the survival of Loans: A case study of banks X and Y*. PhD thesis, Makerere University, 2010.

[237] David C Parkes and Michael P Wellman. Economic reasoning and artificial intelligence. *Science*, 349(6245):267–272, 2015.

[238] Frank Pasquale. *The Black Box Society: The Secret Algorithms that Control Money and Information*. Harvard University Press, 2015.

[239] Kasper Roszbach. Bank lending policy, credit scoring, and the survival of loans. *Review of Economics and Statistics*, 86(4):946–958, 2004.

[240] Jonathan JJM Seddon and Wendy L Currie. A model for unpacking big data analytics in high-frequency trading. *Journal of Business Research*, 70:300–307, 2017.

[241] Robert J Shiller. *Narrative economics: How stories go viral and drive major economic events*. Princeton University Press, 2020.

[242] Chih-Fong Tsai and Jhen-Wei Wu. Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert systems with applications*, 34(4):2639–2649, 2008.

[243] Michael P. Wellman, Peter R. Wurman, Kevin O'Malley, Roshan Bangera, Daniel Reeves, William E Walsh, et al. Designing the market game for a trading agent competition. *IEEE Internet Computing*, 5(2):43–51, 2001.

[244] Alexandre Aronne, Aureliano Bressan, and Haroldo Guimarães Brasil. *Mensuração e Gerenciamento de Riscos Corporativos: Aplicações de Cash Flow at Risk e Real Options*. Saint Paul Editora, 2021.

[245] Rogério Silva Nacif. *Operações Eficientes, Empresas Rentáveis: Melhorando os Resultados Financeiros por Meio da Gestão de Operações*. Aquila, 2021.

[246] Carlo Appugliese, Paco Nathan, and William S Roberts. *Agile AI: A Practical Guide to Building AI Applications and Teams*. O'Reilly, 2020.

[247] Kenneth Appel, Wolfgang Haken, and John Koch. Every planar map is four colorable. part ii: Reducibility. *Illinois Journal of Mathematics*, 21(3):491–567, 1977.

[248] Kenneth Appel and Wolfgang Haken. Every planar map is four colorable. *Bulletin of the American mathematical Society*, 82(5):711–712, 1976.

[249] Per Bak, Kan Chen, and Michael Creutz. Self-organized criticality in the game of life. *Nature*, 342(6251):780–782, 1989.

[250] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

[251] Roger Bemelmans, Gert Jan Gelderblom, Pieter Jonker, and Luc De Witte. Socially assistive robots in elderly care: a systematic review into effects and effectiveness. *Journal of the American Medical Directors Association*, 13(2):114–120, 2012.

[252] Andrew Berdahl, Colin J Torney, Christos C Ioannou, Jolyon J Faria, and Iain D Couzin. Emergent sensing of complex environments by mobile animal groups. *Science*, 339(6119):574–576, 2013.

[253] Ana Berdasco, Gustavo López, Ignacio Diaz, Luis Quesada, and Luis A Guerrero. User experience comparison of intelligent personal assistants: Alexa, google assistant, siri and cortana. *Multidisciplinary Digital Publishing Institute Proceedings*, 31(1):51, 2019.

[254] Alessandro Bessi and Emilio Ferrara. Social bots distort the 2016 us presidential election online discussion. *First monday*, 21(11-7), 2016.

[255] Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin. Heads-up limit hold'em poker is solved. *Communications of the ACM*, 60(11):81–88, 2017.

[256] Murray Campbell, A Joseph Hoane Jr, and Feng-hsiung Hsu. Deep blue. *Artificial intelligence*, 134(1-2):57–83, 2002.

[257] Juan Miguel Carrascosa, Jakub Mikians, Ruben Cuevas, Vijay Erramilli, and Nikolaos Laoutaris. I always feel like somebody's watching me: measuring online behavioural advertising. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*, pages 1–13, 2015.

[258] Maurice Chiodo and Toby Clifton. The importance of ethics in mathematics. *European Mathematical Society Magazine*, 114:34–37, 2019.

[259] BACEN. LIFT Challenge, 2022. Acessed in 03/09/2022.

[260] Wasifa Chowdhury. *Employing neural hierarchical model with pointer generator networks for abstractive text summarization*. PhD thesis, Simon Frazer University: School of Computing Science, 2019.

[261] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

[262] Jack Clark and Ray Perrault. Introduction to the AI index report 2022. Technical report, Stanford University, 2022.

[263] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv preprint arXiv:1408.6491*, 2014.

[264] Yue Deng, Feng Bao, Youyong Kong, Zhiquan Ren, and Qionghai Dai. Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems*, 28(3):653–664, 2016.

[265] Pedro Domingos. *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Basic Books, 2015.

[266] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[267] Sebastian Elbaum and John C Munson. Software black box: an alternative mechanism for failure analysis. In *Proceedings 11th International Symposium on Software Reliability Engineering. ISSRE 2000*, pages 365–376. IEEE, 2000.

[268] Giuliano Da Empoli. *Os engenheiros do caos*. Vestígio, Belo Horizonte, 1 edition, 2019.

[269] Amir Globerson and Sam Roweis. Nightmare at test time: robust learning by feature deletion. In *Proceedings of the 23rd international conference on Machine learning*, pages 353–360, 2006.

[270] Dennis R Grossi. Aviation Recorder Overview. In *International Symposium On Transportation Recorders*, page 12, 2006.

[271] Jose Hernandez-Orallo. Beyond the turing test. *Journal of Logic, Language and Information*, 9(4):447–466, 2000.

[272] John Kay and Mervyn King. *Radical uncertainty: Decision-making beyond the numbers*. WW Norton & Company, 2020.

[273] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. Identifying unknown unknowns in the open world: Representations and policies for guided exploration. In *Thirty-first aaai conference on artificial intelligence*, 2017.

[274] Tian-Shyug Lee and I-Fei Chen. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with applications*, 28(4):743–752, 2005.

[275] Joel Z Leibo, Cyprien de Masson d'Autume, Daniel Zoran, David Amos, Charles Beattie, Keith Anderson, Antonio García Castañeda, Manuel Sanchez, Simon Green, Audrunas Gruslys, et al. Psychlab: a psychology laboratory for deep reinforcement learning agents. *arXiv preprint arXiv:1801.08116*, 2018.

[276] Nancy G. Leveson. *Engineering a Safer World: Systems Thinking Applied to Safety*. The MIT Press, 2011.

[277] C Dianne Martin. The myth of the awesome thinking machine. *Communications of the ACM*, 36(4):120–133, 1993.

[278] Simone Natale et al. *Deceitful media: Artificial intelligence and social life after the Turing test*. Oxford University Press, USA, 2021.

[279] Olfa Nasraoui and Patrick Shafto. Human-Algorithm Interaction Biases in the Big Data Cycle: A Markov Chain Iterated Learning Framework. *arXiv*, 2016.

[280] Randolph M Nesse. Tinbergen's four questions, organized: a response to bateson and laland. *Trends in Ecology & Evolution*, 28(12):681–82, 2013.

[281] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[282] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York, first edition, 2018.

[283] David L. Poole and Alan K. Mackworth. *Artificial Intelligence: foundations of computational agents*. Cambridge University Press, second edition, 2017.

[284] David L Poole and Alan K Mackworth. *Artificial Intelligence: foundations of computational agents*. Cambridge University Press, 2010.

[285] Stuart Russel and Peter Norvig. *Artificial Intelligence*. Prentice Hall, New York, 3 edition, 2010.

[286] Jonathan Schaeffer, Neil Burch, Yngvi Bjornsson, Akihiro Kishimoto, Martin Muller, Robert Lake, Paul Lu, and Steve Sutphen. Checkers is solved. *science*, 317(5844):1518–1522, 2007.

[287] Greg Siegel. *Forensic media: Reconstructing accidents in accelerated modernity*. Duke University Press, 2014.

[288] Robert Skidelsky. *Information retrieval and hypertext*. Yale University Press, 2020.

[289] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[290] Megha Srivastava, Hoda Heidari, and Andreas Krause. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2459–2468, 2019.

[291] Venkatramanan S Subrahmanian, Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. The darpa twitter bot challenge. *Computer*, 49(6):38–46, 2016.

[292] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, pages 1–10, 2014.

[293] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, pages 1–22, 2018.

[294] Milena Tsvetkova, Taha Yasseri, Eric T. Meyer, J. Brian Pickering, Vegard Engen, Paul Walland, Marika Lüders, Asbjørn Følstad, and George Bravos. Understanding Human-Machine Networks. *ACM Computing Surveys*, 50(1):1–35, 2018.

[295] Milena Tsvetkova, Ruth García-Gavilanes, Luciano Floridi, and Taha Yasseri. Even good bots fight: The case of wikipedia. *PloS one*, 12(2):e0171774, 2017.

[296] Alan M Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, 1950.
[297] Koen Van De Sande, Theo Gevers, and Cees Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2009.
[298] Carissa Véliz. *Privacy is power*. Melville House, 2021.
[299] Xingyu Xing, Wei Meng, Dan Doozan, Alex C Snoeren, Nick Feamster, and Wenke Lee. Take this personally: Pollution attacks on personalized services. In *22nd USENIX Security Symposium (USENIX Security 13)*, pages 671–686, 2013.
[300] Yu Yao and Ella Atkins. The smart black box: A value-driven high-bandwidth automotive event data recorder. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1484–1496, 2020.
[301] Quan-shi Zhang and Song-Chun Zhu. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1):27–39, 2018.
[302] Zhiming Zhou, Han Cai, Shu Rong, Yuxuan Song, Kan Ren, Weinan Zhang, Yong Yu, and Jun Wang. Activation maximization generative adversarial nets. *arXiv preprint arXiv:1703.02000*, 2017.
[303] Susmit Jha, Tuhin Sahai, Vasumathi Raman, Alessandro Pinto, and Michael Francis. Explaining ai decisions using efficient methods for learning sparse boolean formulae. *Journal of Automated Reasoning*, 63(4):1055–1075, 2019.
[304] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. Dynabench: Rethinking benchmarking in nlp. *arXiv preprint arXiv:2104.14337*, 2021.
[305] Peter Marc Deisenroth, A. Aldo Faisal, and Cheng Soom Ong. *MATHEMATICS FOR MACHINE LEARNING*. Cambridge University Press, 1 edition, 2020.
[306] Patrícia Gomes Rêgo de Almeida. Regulação da Inteligência Artificial: Ação Coletiva que Requer Governança. *Computação Brasil*, 7:23–26, 2022.
[307] Margaret Boden, Joanna Bryson, Darwin Caldwell, Kerstin Dautenhahn, Lilian Edwards, Sarah Kember, Paul Newman, Vivienne Parry, Geoff Pegman, Tom Rodden, Tom Sorrell, Mick Wallis, Blay Whitby, and Alan Winfield. Principles of robotics: regulating robots in the real world. *Connection Science*, 29(2):124–129, 2017.
[308] Kevin Bonsor and Nathan Chandler. How black boxes work. *HowStuffWorks, June*, 13, 2001.
[309] Fernanda Braganca and Renata Braga. Os Desafios da Regulamentação Jurídica da Inteligência Artificcial no Brasil. *Computação Brasil*, 7:19–22, 2022.
[310] Lindell Bromham, Russell Dinnage, and Xia Hua. Interdisciplinary research has consistently lower funding success. *Nature*, 534(7609):684–687, 2016.
[311] Joanna J Bryson, Mihailis E Diamantis, and Thomas D Grant. Of, for, and by the people: the legal lacuna of synthetic persons. *Artificial Intelligence and Law*, 25(3):273–291, 2017.
[312] Roger Clarke. Regulatory alternatives for ai. *Computer Law & Security Review*, 35(4):398–409, 2019.
[313] Europe Commission. Coordinated Plan on Artificial Intelligence 2021 Review. Technical report, European Commission, Brussels, 2021.
[314] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.
[315] Iain D Couzin, Christos C Ioannou, Güven Demirel, Thilo Gross, Colin J Torney, Andrew Hartnett, Larissa Conradt, Simon A Levin, and Naomi E Leonard. Uninformed

individuals promote democratic consensus in animal groups. *science*, 334(6062):1578–1580, 2011.

[316] Kate Crawford, Meredith Whittaker, Madeleine Clare Elish, Solon Barocas, Aaron Plasek, and Kadija Ferryman. The ai now report. *The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term*, 2016.
O paper analisou a IA em relação a quatro temas-chave: Saúde, Trabalho, Desigualdade e Ética. E oferece oito (8) recomendações como passos práticos que os interessados envolvidos em vários pontos na produção, uso, governança e avaliação dos sistemas de IA poderiam tomar para enfrentar os desafios e oportunidades de curto prazo criados pela rápida implementação de IA através dos domínios social e econômico.

[317] Kenneth Cukier, Viktor Mayer-Schönberger, and Francis de Véricourt. *Framers: Human advantage in an age of technology and turmoil*. Penguin, 2022.

[318] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, pages 1–14, 2016.

[319] EPI. Algorithmic transparency: End secret profiling. Technical report, Electronic Privacy Information Center, 2015.

[320] Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

[321] Gregory Falco, Ben Shneiderman, Julia Badger, Ryan Carrier, Anton Dahbura, David Danks, Martin Eling, Alwyn Goodloe, Jerry Gupta, Christopher Hart, et al. Governing ai safety through independent audits. *Nature Machine Intelligence*, 3(7):566–571, 2021.

[322] Hanming Fang and Andrea Moro. Theories of statistical discrimination and affirmative action: A survey. In *Handbook of Social Economics*, volume 1, pages 133–200. Elsevier B.V., 2011.

[323] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. *Berkman Klein Center Research Publication*, 2020-1, 2020.

[324] Ana Frazão. Discriminação Algoritmica: a relação entre homens e máquinas. Coluna Jota, junho 2021. Trabalho dividido em treze partes.

[325] Future of Live Institute. Autonomous Weapons: An Open Letter from AI and Robotics Reserarchers. `https://futureoflife.org/2016/02/09/open-letter-autonomous-weapons-ai-robotics/?cn-reloaded=1&cn-reloaded=1`, July 2018.

[326] IEEE. Ethically Aligned Design: Version 2 - For Public Discussion. *IEEE Standards*, pages 1–263, 2017.

[327] CEI. Our vision. Technical report, Council on ExtendedIntelligence, 2022.

[328] Lucas D. Introna and Helen Nissenbaum. Shaping the web: Why the politics of search engines matters. *Computer Ethics*, pages 157–173, 2017.

[329] Joosr. *A Joosr Guide to... Weapons of Math Destruction by Cathy O'Neil: How Big Data Increases Inequality and Threatens Democracy*. Broadway books, 2016.

[330] Daniel Kahneman, Olivier Sibony, and CR Sunstein. *Noise*. HarperCollins UK, 2022.

[331] Daniel Kahneman, AM Rosenfield, L Gandhi, and T Blaser. Noise: How to overcome the high. *Havard Business Review*, 2016. Available in https://hbr.org/2016/10/noise.

[332] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[333] Pratyusha Kalluri et al. Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, 583:169–169, 2020.

[334] Krishna M. Kavi. Beyond the Black Box. *IEE Spectrum*, 47(8):46—-51, 2021.

[335] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T Hancock, and Michael S Bernstein. Conceptual metaphors impact perceptions of human-ai collaboration. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–26, 2020.

[336] Mervyn King and John Kay. *Radical Uncertainty: Decision-making for an unknowable future*. Hachette UK, 2020.

[337] Nicole C Krämer, Astrid von der Pütten, and Sabrina Eimler. Human-agent and human-robot interaction theory: Similarities to and differences from human-human interaction. In *Human-computer interaction: The agency perspective*, pages 215–240. Springer, 2012.

[338] Samantha Krening and Karen M Feigh. Interaction Algorithm Effect on Human Experience. *ACMTrans. Human-Robot Interact*, 7(2):22, 2018.

[339] Armin Krishnan. *Killer robots: legality and ethicality of autonomous weapons*. Routledge, 2016.

[340] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

[341] Heidi Ledford. Team science. *Nature*, 525(7569):308–311, 2015.

[342] Joseph CR Licklider. Man-computer symbiosis. *IRE transactions on human factors in electronics*, pages 4–11, 1960.

[343] Claudia Bauzer Medeiros. Dados, Algoritmos, Máquinas e Pessoas. *Computação Brasil*, 7:11–14, 2022.

[344] David A Mindell. *Our robots, ourselves: Robotics and the myths of autonomy*. Viking, 2015.

[345] Sendhil Mullainathan. Biased algorithms are easier to fix than biased people. *The New York Times*, 2019.

[346] NSCAI. National Security Commission on Artificial Intelligence - Interim Report. *National Security Commission on Artificial Intelligence Report*, pages 1–101, 2019.

[347] Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway books, 2016.

[348] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[349] Cristina Godoy Bernardo de Oliveira, João Paulo Cândia Veiga, and Fabio G. Cozman. Regulação da Inteligência Artificial: Qual o Modelo Adotar. *Computação Brasil*, 7:28–31, 2022.

[350] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, 30(3):286–297, 2000.

[351] Frank Pasquale. *New laws of robotics: defending human expertise in the age of AI*. Belknap Press, 2020.

[352] Komal Patel. Testing the Limits of the First Amendment: How a CFAA Prohibition on Online Antidiscrimination Testing Infringes on Protected Speech Activity. *SSRN Electronic Journal*, pages 1–46, 2017.

[353] Walt L Perry. *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation, 2013.

[354] Sundar Pichai. AI at Google: our principles. *2018-07-07*, page 1, 2018.

[355] Tony J Prescott and Julie M Robillard. Are friends electric? the benefits and risks of human-robot relationships. *Iscience*, 24(1):101993, 2021.

[356] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al. Machine behaviour. *Nature*, 568(7753):477–486, 2019.
Trata do comportamento dos sistemas de AI, cuja compreensão é essencial para nossa capacidade de controlar suas ações, colher seus benefícios e minimizar seus danos. Em síntese, entender o comportamento dos sistemas de AI, para maximizar seus benefícios e minimizar os danos em relação à humanidade.

[357] Byron Reeves and Clifford Nass. How people treat computers, television, and new media like real people and places, 1996.

[358] Tahira Reid and James Gibert. Inclusion in human–machine interactions. *Science*, 375(6577):149–150, 2022.

[359] Lionel P Robert. The Growing Problem of Humanizing Robots. *International Robotics & Automation Journal*, 3(1):1–2, 2017.

[360] Margaret E Roberts. Censored. In *Censored*. Princeton University Press, 2018.

[361] W Teed Rockwell. Algorithms and stories. *Human Affairs*, 23(4):633–644, 2013.

[362] Heather M Roff. The strategic robot problem: Lethal autonomous weapons in war. *Journal of Military Ethics*, 13(3):211–227, 2014.

[363] Jário Santos, Ig Bittencourt, Marcelo Reis, Geiser Chalco, and Seiji Isotani. Two billion registered students affected by stereotyped educational environments: an analysis of gender-based color bias. *HUMANITIES AND SOCIAL SCIENCES COMMUNICATIONS |*, 9(249):1—-16, 2022.

[364] Daniel Schiff, Aladdin Ayesh, Laura Musikanski, and John C Havens. Ieee 7010: A new standard for assessing the well-being implications of artificial intelligence. In *2020 IEEE international conference on systems, man, and cybernetics (SMC)*, pages 2746–2753. IEEE, 2020.

[365] Helen Sharp, Yvonne Rogers, and Jennifer Preece. *Interaction Design: Beyond Human-Computer Interaction*. John Wiley & Sons Inc, fifth edition, 2019.

[366] Ben Shneiderman. Design lessons from ai's two grand goals: Human emulation and useful applications. *IEEE Transactions on Technology and Society*, 1(2):73–82, 2020.

[367] Ben Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020.

[368] Ben Shneiderman. The dangers of faulty, biased, or malicious algorithms requires independent oversight. *Proceedings of the National Academy of Sciences*, 113(48):13538–13540, 2016.

[369] Ben Shneiderman, Catherine Plaisant, Maxine S Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. *Designing the user interface: strategies for effective human-computer interaction*. Pearson, 2016.

[370] Ben Shneiderman. Human responsibility for autonomous agents. *IEEE intelligent systems*, 22(2):60–61, 2007.

[371] Ben Shneiderman and Pattie Maes. Direct manipulation vs. interface agents. *interactions*, 4(6):42–61, 1997.

[372] Ben Shneiderman. The future of interactive systems and the emergence of direct manipulation. *Behaviour & Information Technology*, 1(3):237–256, 1982.

[373] Ben Shneiderman. Direct manipulation: A step beyond programming languages. In *Proceedings of the Joint Conference on Easier and More Productive Use of Computer Systems.(Part-II): Human Interface and the User Interface-Volume 1981*, page 143, 1981.

[374] Marlo Souza. Tecnologias da Linguagem, Ética em IA e Regulamentação. *Computação Brasil*, 7:32–35, 2022.

[375] Peter Stone, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, Shivaram Kalyanakrishnan, Ece Kamar, Sarit Kraus, Kevin Leyton-Brown, David Parkes, William Press, AnnaLee (Anno) Saxenian, Julie Shah, Milind Tambe, and Astro Teller. Artificial Intelligence and Life in 2030: the one hundred year study on artificial intelligence. Technical report, Stanford University, September 2016.

[376] Megan K Strait, Cynthia Aguillon, Virginia Contreras, and Noemi Garcia. The public's perception of humanlike robots: Online social commentary reflects an appearance-based uncanny valley, a general fear of a "technology takeover", and the unabashed sexualization of female-gendered robots. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 1418–1423. IEEE, 2017.

[377] Barry Strauch. Ironies of automation: Still unresolved after all these years. *IEEE Transactions on Human-Machine Systems*, 48(5):419–433, 2017.

[378] Wenlong Sun, Olfa Nasraoui, and Patrick Shafto. *Evolution and impact of bias in human and machine learning algorithm interaction*, volume 15. Public Library of Science San Francisco, CA USA, 2020.

[379] Latanya Sweeney. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54, 2013.

[380] Niko Tinbergen. On aims and methods of ethology. *Animal Biology*, 55(4):297–321, 2005.

[381] Ufuk Topcu, Nadya Bliss, Nancy Cooke, Missy Cummings, Ashley Llorens, Howard Shrobe, and Lenore Zuck. Assured autonomy: Path toward living with autonomous systems we can trust. *arXiv preprint arXiv:2010.14443*, 2020.

[382] Zeynep Tufekci. Youtube, the great radicalizer. *The New York Times*, 10(3):2018, 2018.

[383] Zeynep Tufekci. Engineering the public: Big data, surveillance and computational politics. *First Monday*, 2014.

[384] UN. Report of the Working Group on Internet Governance. Technical report, United Nations, 2005.

[385] University of Montreal. Montréal Declaration for a Responsible Development of Artificial Intelligence. Technical report, University of Montreal, 2018.

[386] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.

[387] Paul Voosen. The ai detectives, 2017.

[388] Sara Wachter-Boettcher. *Technically wrong: Sexist apps, biased algorithms, and other threats of toxic tech*. WW Norton & Company, 2017.

[389] Chathurika S Wickramasinghe, Daniel L Marino, Javier Grandio, and Milos Manic. Trustworthy ai development guidelines for human system interaction. In *2020 13th International Conference on Human System Interaction (HSI)*, pages 130–136. IEEE, 2020.

[390] Christine T. Wolf and Jeanette L. Blomberg. Evaluating the promise of human-algorithm collaborations in everyday work practices. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

[391] David D Woods, James Tittle, Magnus Feil, and Axel Roesler. Envisioning human-robot coordination in future operations. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(2):210–218, 2004.

[392] Wei Xu. Toward human-centered ai: a perspective from human-computer interaction. *interactions*, 26(4):42–46, 2019.

[393] Sergio Amadeu Silveira. Governo dos algoritmos. *Revista de Políticas Públicas*, 21(1):267–281, 2017.

[394] Can Yavuz. *Machine Bias Artificial Intelligence and Discrimination*. PhD thesis, Lund University, 2019.

[395] Shunyuan Zhang, Nitin Mehta, Param Vir Singh, and Kannan Srinivasan. Can an ai algorithm mitigate racial economic inequality? an analysis in the context of airbnb. *An Analysis in the Context of Airbnb (January 21, 2021). Rotman School of Management Working Paper*, 2021.

[396] Ignas Kalpokas. *Algorithmic Governance*, volume 9. Palgrave Macmillan, 2019.

[397] Meredith Broussard. *Artificial unintelligence: How computers misunderstand the world*. MIT Press, 2018.

[398] César A Hidalgo, Diana Orghian, Jordi Albo Canals, Filipa De Almeida, and Natalia Martín. *How humans judge machines*. MIT Press, 2021.

[399] Alexis Lambert, Nahal Norouzi, Gerd Bruder, and Gregory Welch. A systematic review of ten years of research on human interaction with social robots. *International Journal of Human–Computer Interaction*, 36(19):1804–1817, 2020.

[400] Yi Mou, Changqian Shi, Tianyu Shen, and Kun Xu. A systematic review of the personality of robot: Mapping its conceptualization, operationalization, contextualization and effects. *International Journal of Human–Computer Interaction*, 36(6):591–605, 2020.

[401] Sarah Sebo, Brett Stoll, Brian Scassellati, and Malte F Jung. Robots in groups and teams: a literature review. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–36, 2020.

[402] Weiyu Wang and Keng Siau. Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: A review and research agenda. *Journal of Database Management (JDM)*, 30(1):61–79, 2019.

[403] Cigdem BAŞFIRINCI and Zuhal ÇİLİNGİR. Anthropomorphism and advertising effectiveness: Moderating roles of product involvement and the type of consumer need. *Journal of Social and Administrative Sciences*, 2(3):108–131, 2015.

[404] Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot, editors. *Media Technologies: Essays on Communication, Materiality, and Society*. The MIT Press, 2014.

[405] Mark Granovetter and Roland Soong. Threshold models of diffusion and collective behavior. *Journal of Mathematical sociology*, 9:165–179, 1983.

[406] Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83:1420–1443, 1978.

[407] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1914–1933, 2017.

[408] Leila Hudson, Colin S Owens, and Matt Flannes. Drone warfare: Blowback from the new american way of war. *Middle East Policy*, 18:122–132, 2011.

[409] Colm Kearns, Gary Sinclair, Jack Black, Mark Doidge, Thomas Fletcher, Daniel Kilvington, Katie Liston, Theo Lynn, and Pierangelo Rosati. A scoping review of research on online hate and sport. *Communication & Sport*, pages 1–29, 2022.

[410] Peter M Krafft, Michael Macy, and Alex" Sandy" Pentland. Bots as virtual confederates: design and ethics. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 183–190, 2017.

[411] Thomas S. Kuhn. *A Estrutura das Revoluções Científicas*. Perspectiva, São Paulo, 1 edition, 1996.

[412] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Computational social science. *Science*, 323(5915):721–723, 2009.

[413] Yuhua Liang and Seungcheol Austin Lee. Fear of autonomous robots and artificial intelligence: Evidence from national representative data with probability sampling. *International Journal of Social Robotics*, 9(3):379–384, 2017.

[414] C Dianne Martin. Eniac: press conference that shook the world. *IEEE Technology and Society Magazine*, 14(4):3–10, 1995.

[415] Neil McBride. Robot enhanced therapy for autistic children: An ethical analysis. *IEEE Technology and Society Magazine*, 39(1):51–60, 2020.

[416] Bjarke Mønsted, Piotr Sapieżyński, Emilio Ferrara, and Sune Lehmann. Evidence of complex contagion of information in social media: An experiment using twitter bots. *PloS one*, 12(9):e0184148, 2017.

[417] Lewis Mumfordt. *Technics and Civilization*. University of Chicago Press, Chicago, 1934.

[418] Kathleen Richardson, Mark Coeckelbergh, Kutoma Wakunuma, Erik Billing, Tom Ziemke, Pablo Gomez, Bram Vanderborght, and Tony Belpaeme. Robot enhanced therapy for children with autism (dream): A social model of autism. *IEEE Technology and society magazine*, 37(1):30–39, 2018.

[419] Michael P Wellman and Uday Rajan. Ethical issues for autonomous trading agents. *Minds and Machines*, 27(4):609–624, 2017.

[420] Marc Wiedermann, E Keith Smith, Jobst Heitzig, and Jonathan F Donges. A network-based microfoundation of granovetter's threshold model for social tipping. *Scientific reports*, 10(1):1–10, 2020.

[421] Alan FT Winfield and Marina Jirotka. The case for an ethical black box. In *Annual Conference Towards Autonomous Robotic Systems*, pages 262–273. Springer, 2017.

[422] Alan FT Winfield and Marina Jirotka. Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133):20180085, 2018.

[423] Néstor García Canclini. *Ciudadanos reemplazados por algoritmos*. Calas, 2019.

[424] Joseph Weizenbaum. *Computer Power and Human Reason: From Judgement to Calculation*. W. H. Freeman and Company, 1976.