

Defining AI in Policy versus Practice

P. M. Krafft*
Oxford Internet Institute
University of Oxford
p.krafft@oii.ox.ac.uk

Meg Young
Information School
University of Washington
megyoung@uw.edu

Michael Katell
Information School
University of Washington
mkatell@uw.edu

Karen Huang
Harvard University
karenhuang@g.harvard.edu

Ghislain Bugingo
Information School
University of Washington

ABSTRACT

Recent concern about harms of information technologies motivate consideration of regulatory action to forestall or constrain certain developments in the field of artificial intelligence (AI). However, definitional ambiguity hampers the possibility of conversation about this urgent topic of public concern. Legal and regulatory interventions require agreed-upon definitions, but consensus around a definition of AI has been elusive, especially in policy conversations. With an eye towards practical working definitions and a broader understanding of positions on these issues, we survey experts and review published policy documents to examine researcher and policy-maker conceptions of AI. We find that while AI researchers favor definitions of AI that emphasize technical functionality, policy-makers instead use definitions that compare systems to human thinking and behavior. We point out that definitions adhering closely to the functionality of AI systems are more inclusive of technologies in use today, whereas definitions that emphasize human-like capabilities are most applicable to hypothetical future technologies. As a result of this gap, ethical and regulatory efforts may overemphasize concern about future technologies at the expense of pressing issues with existing deployed technologies.

CCS CONCEPTS

• **Social and professional topics** → **Computing / technology policy**; • **Computing methodologies** → *Artificial intelligence*.

KEYWORDS

definitions; policy; artificial intelligence; sociotechnical imaginaries

ACM Reference Format:

P. M. Krafft, Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo. 2020. Defining AI in Policy versus Practice. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*, February 7–8, 2020, New York, NY, USA. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3375627.3375835>

*Work conducted at the University of Washington.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

AIES '20, February 7–8, 2020, New York, NY, USA

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-1-4503-7110-0/20/02...\$15.00
<https://doi.org/10.1145/3375627.3375835>

1 INTRODUCTION

As computational systems have come to play an increasing role in making predictions about people, the downstream social consequences of artificial intelligence (AI) have garnered growing public attention. In the short term, evidence indicates that machine learning (ML) algorithms could contribute to oppression and discrimination due to historical legacies of injustice reflected in training data [2, 7, 13], directly to economic inequality through job displacement [14], or through a failure to reflectively account for who benefits and who does not from the decision to use a particular system [18]. In the long term, some believe AI technologies pose an existential risk to humanity by altering the scope of human agency and self-determination [24], or by the creation of autonomous weapons [3]. For many researchers in machine learning, the answer to these challenges has been technical, and there has been a growth of work to promote computational approaches to fairness, accountability, and transparency in machine learning (e.g., [5, 21]). For others, these challenges may require regulatory intervention. In the regulatory line, widespread public concerns regarding the social impacts of AI have led a growing number of organizations to create policy and ethical recommendations for AI governance [23].

Here, we examine the regulatory approach and the extent to which current policy efforts are in meaningful conversation with existing AI and machine learning research, beginning with the simplest matter of definitions. Given that lawmakers are not *prima facie* technologists or AI researchers, it is important to understand policymakers' operational definitions of AI on the path to effective governance choices. Policymakers' understandings of AI may differ from those of AI researchers. For example, recent findings on municipal technology policy found that government employees did not define Automated License Plate Readers (ALPR) or Booking Photo Comparison Software, which rely on optical character recognition and facial recognition, as AI or machine learning systems [36]. Where policymakers do not have a clear definition of AI for regulatory purposes [19] bureaucrats do not know which systems fall under new laws in the implementation phase. Conceptual clarity on this issue would empower a new generation of algorithmic oversight regulation.

A *policy-facing definition* of AI is one that facilitates policy implementation. Definitional challenges of automation persist across domains. In the case of autonomous weapons, a major barrier to consensus in international discussions has been lack of agreement over the definition of an autonomous weapon [11, 25]. Furthermore, policy documents often use definitions of AI that are difficult to

apply from the standpoint of policy implementation. For example, the AI4People Scientific Committee defines AI as “a resource [that] can hugely enhance human agency” [17]. As another example, the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems defines AI as “autonomous and intelligent technical systems [that] are specifically designed to reduce the necessity for human intervention in our day-to-day lives” [4]. While these definitions are conceptually illuminating in highlighting the role of humans in AI, they are (i) too ambiguous to usefully apply in regulatory approaches to governance, and (ii) tend to misapprehend AI’s current capabilities. Policymakers concerned about the disparate impact of ML applications thus risk overlooking currently deployed technology in favor of next-generation or speculative systems.

This lack of a policy-facing definitions, along with a possible disconnect between policymakers’ and AI researchers’ conceptions of AI, motivate the current study. How do policy documents and AI researchers conceptualize AI technologies? How may these conceptualizations differ? While previous work has hinted at answers to these questions, no systematic evaluation has been conducted. We employ a mixed methods social scientific approach to address these questions. Drawing upon on Russell and Norvig’s typology of AI definitions [26], we first use this typology to analyze results from a series of surveys of researchers in AI to understand how researchers define AI. We then conduct a document analysis of major published policy documents in order to understand policymakers’ understanding of AI through a close analysis of the definitions those documents use. Our use of a single well-established typology of AI definitions allows us to systematically compare disparate survey and document data sources to achieve a meaningful understanding of how published policy documents’ definitions of AI compare to AI researchers’ definitions of AI.

2 PREVIOUS WORK

Researchers in AI have long recognized the lack of definitional consensus in the field. According to Agre [1], the lack of a definition was generative: “Curiously, the lack of a precise, universally accepted definition of AI probably has helped the field to grow, blossom, and advance.” Even as the field has disagreed widely in practice, in their foundational textbook, “Artificial Intelligence: A Modern Approach,” Russell and Norvig [26] examine the goals of the field by considering how AI is defined in eight other textbooks published between 1978-1993, finding that definitions can be classified into four primary approaches. Namely, AI as a field is concerned with building systems that (i) think like humans, (ii) act like humans, (iii) think rationally, and (iv) act rationally. (See Table 1.) However, according to Sweeney [32], the lack of alignment between different conceptions of AI poses a risk to the field. In order to advance a critique that the term AI suffers from “too many different and sometimes conflicted meanings,” Sweeney used this typology to characterize 996 AI research works cited in Russell and Norvig’s textbook based on her sense that a textbook cites works representative of or important to the field. Of these, nearly all (987) fell into one of the latter two approaches, i.e. to create artifacts that pursue rational (a.k.a. “ideal”) thinking and behavior. Sweeney also notes large definitional shifts in the field over time, as well as low proportions of references in common between textbooks

	Behaving	Thinking
Human	Definitions emphasizing meeting human performance or behaving like people do in particular tasks or situations	Definitions emphasizing processing information in ways inspired by human cognition
Ideal	Definitions emphasizing functioning optimally according to a specification of optimality, such as by accomplishing specified tasks or goals	Definitions emphasizing processing information according to idealized rules of inference

Table 1: Russell and Norvig [26] divided definitions of AI according to two dimensions: human versus ideal, and thinking versus behaving. All definitions of AI are classifications of computer programs as “intelligent” or not according to some characteristics. A human thinking definition of AI classifies a machine or a computer program as AI as being able to imitate human cognition, i.e. if it is able to imitate the way people think, either in a particular task or in a range of situations. In contrast, an ideal thinking definition is one that emphasizes AI as reasoning correctly according to mathematical rules given a set of assumptions about the world, such as deductive logic or statistical inference. A human behaving definition of AI emphasizes AI as imitating the way people behave or complete particular tasks—the distinction between human thinking versus behaving is that a human behaving AI might be able to do the same thing a person does but for different reasons, such as using hard-wired rules. Finally, an ideal behaving definition emphasizes AI as correctly accomplishing the tasks or goals it is programmed to accomplish. Ideal behaving AI might “think” (i.e., process information) and behave in unusual or unfamiliar ways, but ways that are effective for is specified tasks or goals.

published only a few years apart. Some definitional variation may proceed from expectations of AI that are relative to the current capabilities of computing. An early AI researcher remarked that “AI is a collective name for problems which we do not yet know how to solve properly by computer” [22]. Similarly, John McCarthy states, “As soon as it works, no one calls it AI anymore” [33]. In recent years, there has been a growing amount of systematic inquiry into non-expert descriptions and understandings of AI. Awareness of AI is growing; Cave et al. [8] find that in a nationally representative survey in the United Kingdom, 85% of respondents had heard of AI. Among a subset of 622 people who provided a description of AI, 261 referred to computers performing tasks like “decision making,” “learning,” or “thinking.” Another 155 respondents described it as synonymous with robots. As the authors note, “This conflation is understandable...[but] could be problematic. Imagining AI as embodied will lend itself to some narratives more than others [such as] worries of gun-toting killer robots rather than the real-world risk of algorithmic bias” [8]. Confirming this finding, other recent work to understand non-experts’ primary concerns about AI has found that popular imaginaries draw on imagery from media [16]. A focus

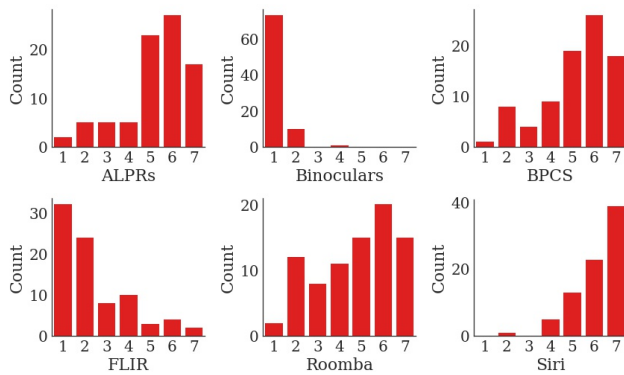


Figure 1: Histograms of AI researcher ratings of technologies as relying on AI. (1 = Strongly disagree, 7 = Strong agree)

on humanoid AI contributes to misinformed public perception and “overshadow[s] issues that are already creating challenges today” [9]. In response to these challenges, scholars argue for changing language used about AI to sharpen its conceptual clarity. Johnson and Verdicchio [20] identify two primary concerns with the public reception of expert discourse on AI. First, the concept of “autonomy” is misapprehended as machines possessing free will and self interest. Second, a discussion highlighting the “system” has obscured the degree to which applications are *sociotechnical*, reflecting a range of human choices, negotiations, and impacts in their design and use. To address these concerns, the authors advocate for re-framing conversations around machine autonomy to foreground human actors and the broader sociotechnical context in which such systems are embedded, an injunction forcefully argued by other critical scholars as well [15]. Other authors, such as Joanna Bryson [6], have offered definitions that begin to synthesize these considerations with classical definitions of AI, conceived of as idealized sensing-acting machines.

3 METHODS

The aims of this study are twofold: To describe the way that AI researchers and policymakers define AI, and to examine possible gaps between these definitions as they relate to implications for policy-making. We employed a multi-method social scientific approach. Our methodological motivation was complementary in combining different types of data [29].

We conducted two surveys of AI and ML researchers in May and July 2019. Our first survey includes data from 98 researchers and our second includes data from 86. We confirmed that our sample has the experience and expertise relevant to our research questions. We also assessed sample bias by confirming that the demographics of our sample match the demographics of the field of AI. Further, we replicated both surveys with an alternative snowball sampling recruitment methodology with 108 and 44 researchers participating. Images of all the stimuli used in these surveys and results from our snowball sample replication are presented in our supplementary materials.¹

¹Full paper with supplementary materials is available online at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3431304.

3.1 AI Researcher Survey

We first conducted two surveys to understand AI researchers’ opinions regarding definitions of AI and views on the social impacts of AI. In the first of these researcher surveys, participants began by rating the extent to which they would classify six particular technologies as AI. In choosing the six particular technologies we presented, we drew on prior work of ours that identified different conceptions of certain surveillance technologies among city stakeholders in Seattle [36]. We chose four surveillance technologies drawn from this previous study, and supplemented those four with a popular virtual assistant and a popular robot. These six technologies were Automated License Plate Readers (ALPRs), binoculars, Booking Photo Comparison Software (BPCS), Forward-Looking Infrared camera (FLIR), Roomba, and Siri. For each technology, we presented images and descriptions, and we asked, “Please indicate your response to the following statement: ALPRs (Automated License Plate Readers) are an artificial intelligence (A.I.)-based technology.” We presented a 7-point Likert scale ranging from “Strongly disagree” to “Strongly agree”. We presented these technologies in randomized order. After these initial questions on classifying AI technologies, we then offered participants the option to define the term AI themselves. Finally, participants provided their opinions regarding particular social impacts of AI. Based on our understanding of the public discourse around AI, we focused on the harms of existential risk, economic inequality, and oppression and discrimination. The format of our questions in this part of the survey was in the form “Please indicate your response to the following statement: The potential for an existential threat to humanity is a relevant issue to consider in the research and development of A.I. technologies.” We used the same 7-point Likert scale as in our technology classification questions.

In our second survey, rather than soliciting classifications of technologies or soliciting definitions of AI, we asked participants to directly selected the characteristics of definitions of AI they identified as most important and directly selected a favored example definition among one definition of each of the four types we used as classes of definitions.

3.1.1 Survey Recruitment. We used multiple recruitment methods for our surveys. In our first survey, we successfully recruited from two major international AI mailing lists, and from AI and ML mailing lists at four U.S. universities highly ranked in AI and ML according to the U.S. News graduate ranking² and one highly ranked university in the U.K. Survey participation was voluntary and uncompensated. In total, 195 people opened a survey link, and 103 participants completed the survey. We use partial data from all participants who consented. In our analysis we subsampled our data to only include participants self-identifying as AI researchers or publishing in at least one of five AI/ML conferences we asked about (NeurIPS, ICML, ICLR, IJCAI, or AAAI). This left a final sample size of 98 AI researchers for our analyses. We collected demographic information to assess the representativeness of the data we analyzed from this survey. We also replicated our results in a snowball sample through social media and the authors’ professional networks.

²<https://www.usnews.com/best-graduate-schools/top-science-schools/artificial-intelligence-rankings>

In our second survey we successfully recruited from two major international AI mailing lists, and one U.S. university mailing list oriented towards computer science that includes alumni and visitors. Survey participation was voluntary and uncompensated. In total, 201 people opened the survey link, and 112 participants completed the survey. We use partial data from all participants who consented. In our analysis we again subsampled our data to only include participants self-identifying as AI researchers or publishing in at least one of five AI/ML conferences we asked about (NeurIPS, ICML, ICLR, IJCAI, or AAAI). This left a final sample size of 86 AI researchers for our analyses. We collected demographic information to assess the representativeness of the data we analyzed from this survey. We also replicated this survey in a snowball sample through social media and the authors' professional networks.

3.1.2 Survey Population Expertise. Of the AI researcher participants in our survey, all but one reported having at some point now or in the past considered themselves AI researchers. 68% reported having extensive experience in AI research or development, and 15% reported having some formal education or work experience relating to AI. 100% reported that they ever read academic publications/scholarly papers related to AI or machine learning. 22% reported having published at NeurIPS, 19% at ICML, 17% at AAAI, 13% at IJCAI, 7% at ICLR, and 23% at other conferences (including UAI, CVPR, AAMAS, ICRA, IROS, and "IEEE"). 78% reported that they considered any of their own work or projects as "very related" to the development of AI, while 18% reported their work or projects as only "somewhat related," and 4% as "a little related."

3.1.3 Selection Bias in our Sample Population. To assess the representativeness of our biased sample, we checked the distribution of self-reported demographic variables. In our first survey, 25% reported Female as their gender, 70% reported Male, 1% entered their own gender identity, and 4% indicated they preferred not to say. Using categories based on the U.S. Census Bureau classification system, and allowing multiple selections, our subsample of AI researchers consisted of 0% identifying as American Indian or Alaska Native, 26% as Asian, 1% Black or African American, 4% Hispanic, 0% Native Hawaiian or Pacific Islander, and 60% White. 4% entered their own race or ethnicity and 5% indicated they preferred not to say. In our second survey, which used a different phrasing of our gender question [31], 21% reported Woman as their gender, 77% Man, and 0% Non-binary. 0% entered their own gender identity and 2% indicated they preferred not to say. Using categories based on the U.S. Census Bureau classification system, and allowing multiple selections, our subsample of AI researchers consisted of 0% identifying as American Indian or Alaska Native, 17% as Asian, 6% Black or African American, 5% Hispanic, 0% Native Hawaiian or Pacific Islander, and 61% White. 3% entered their own race or ethnicity, and 8% indicated they preferred not to say. Authoritative statistics on the demographics of the fields of AI and machine learning were not readily available at the time of writing [34], but existing estimates place the percentage of women publishing in AI and ML at around 15-20%, with the percentage of women working in computing more broadly being last estimated as around 25% [34].³⁴ Estimates of

³<https://jfgagne.ai/talent-2019/>

⁴<https://www.wired.com/story/artificial-intelligence-researchers-gender-imbalance/>

Type	Definition	Source
Human behavior	The science of making computers do things that require intelligence when done by humans	The Alan Turing Institute
Human thought	The imitation of human intelligence processes by machines, especially computer systems.	Deutsche Telekom
Ideal behavior	Systems that display intelligent behaviour by analysing their environment and taking actions—with some degree of autonomy—to, achieve specific goals.	EU Commission
Ideal thought	AI is computer programming that learns and adapts	Google

Table 2: Example definitions of AI sorted by Russel & Norvig typology.

racial and ethnic diversity are less reliable but place the percentage of black researchers in AI potentially as low as less than 1% and only up to roughly 5% working in tech companies more broadly. One estimate places the percentage of Hispanic tech workers at around the same level [34]. We were not able to find counts of other racial or ethnic groups in the field of AI. A Pew Report shows computer workers in the U.S. as mostly White (65%), followed by Asian (19%), Black (7%), and Hispanic (7%).⁵ According to these estimates, the demographics of our survey appear to be relatively representative of the community of AI researchers along these dimensions, which suggests, at least along these dimensions, the bias in our data from non-random sampling is small.

3.2 Document Analysis of Policy Documents

We also analyzed published policy documents in order to explore how policymakers conceptualize AI. We drew upon a comprehensive inventory of available AI ethics and policy documents published between 2017 and 2019 indexed in the prominent international non-profit research organization, Algorithm Watch's, AI Ethics Guidelines Global Inventory.⁶ The inventory contains documents authored by governments, non-governmental organizations, and firms offering guidance on AI governance. Starting with the 83 documents in the collection at the time of data collection in July 2019, we analyzed all English-language documents whose titles mentioned artificial intelligence (40 documents), removing one duplicate document from the analysis. Within each document, we identified definitions of AI using keyword searching, using the stemmed keywords "defin", "artificial intelligence", "AI", and "A.I.". We analyzed the definitions identified through this search using a qualitative coding procedure. Three authors independently conducted coding; one coder works in artificial intelligence, one in science and technology studies, and one in information policy and ethics. Following Sweeney [32], we classify definitions being used by researchers and policy documents according to Russell and Norvig's taxonomy [26] describing four general approaches to defining AI. For each of these categories, we coded the approach that was most pronounced in emphasis in the definition. At least two of the three coders agreed on all but two cases (95%).

⁵<https://www.pewsocialtrends.org/2018/01/09/diversity-in-the-stem-workforce-varies-widely-across-jobs/>

⁶<https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/>

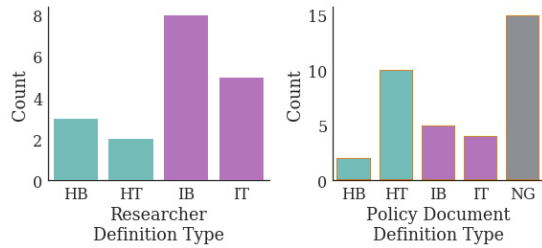


Figure 2: Histogram of types of definitions AI researchers gave in our first survey, which favor ideal (rational) behavior/thinking definitions over human-imitating behavior/thinking, compared to histogram of types of definitions published AI policy document gave, which favor human-type definitions over ideal-type. Codes: I = Ideal (rational), H = Human-imitating, T = Thinking, B = Behaving (following Russell and Norvig), NG = None given.

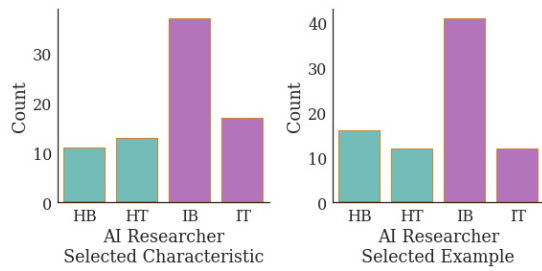


Figure 3: Histogram of types of definitions AI researchers selected as favored characteristics and best example definitions in our second survey. Codes as in Figure 2.

4 RESULTS

4.1 AI Researcher Survey

In our first survey, in which we asked participants to classify technologies as AI or not without providing a definition, we found a large degree of consensus in the classifications of five of the six technologies we presented. Figure 1 shows that AI researchers tend to agree that ALPRs (80% agree at least somewhat), BPCS (74%), and Siri (93%) are AI technologies, while binoculars (0%) and FLIR (11%) are not. The one technology for which there was substantial disagreement was Roombas, with only 60% classifying Roombas as AI. These results indicate that, at least when it comes to classifying particular technologies as AI or not, researchers often more or less agree, even in cases such as ALPRs that have been observed to be contentious for policymakers in prior research.

We also asked an optional question in this first survey for participants to offer their own definition of AI. 18 researchers wrote in definitions, which two authors independently coded according to Russell and Norvig’s typology of human-imitating thinking (HT), human-imitating behavior (HB), ideal (rational) thinking (IT), and ideal behavior (IB). Initial agreement across the two coders was 56%, and the two coders came to consensus where there were disagreements. Consistent with a 2003 analysis of types of AI in the

published literature at the time [32], Figure 2 shows that AI researchers tend to favor ideal thinking/behavior in their definitions of AI (72% used ideal). In other words, AI researchers who provided definitions in our first survey favored definitions of AI that emphasize mathematical problem specification and system functionality over definitions that compare AI to humans.

Our second survey (results shown in Figure 3) was designed to provide a more direct test of this finding from the first survey. Here we asked only for demographics and two fine-grained questions about how to define AI. We presented four example definitions that typify each definition category (HT, HB, IT, IB) and asked participants to select the best definition. We also directly asked which among the categories human-like thinking, human-like behavior, ideal/rational thinking, or ideal/rational behavior the participants judged to be most important to characterizing AI. This simplified survey design qualitatively replicated the result of our first survey. We found that 65% of AI researchers in this survey preferred ideal-type example definitions and 66% of AI researchers selected the ideal-type categories of definitions as most important for characterizing AI.

4.2 Policy Document Analysis

Since the population of AI policymakers is less readily accessible than the population of AI researchers, we rely on content analysis of published policy documents rather than an interview to characterize the types of definitions of AI that tend to be used in policy-making. The documents we analyzed are published by governments, non-governmental organizations, and companies actively engaged in AI ethics and regulation (e.g. AI Now, EU Commission, Microsoft, Open AI, etc.). AI policy and ethics documents are the documents that policymakers consult and produce while crafting regulation and are more readily accessible than the broadly distributed set of people drafting them. This comparison allows us to analyze what published policy documents say about AI to what AI researchers think about AI.

We find that, in contrast to AI researchers, published policy documents tend to use definitions of AI that emphasize comparison to human thought or behavior (57% of definitions place this emphasis). For example, the Science and Technology Committee in the UK Parliament House of Commons defines AI as “a set of statistical tools and algorithms that combine to form, in part, intelligent software... to simulate elements of human behaviour such as learning, reasoning and classification” [27]. Other definitions mention that AI is “inspired by the human mind” [12], “capable of learning from experience, much as people do” [30], or “seek to simulate human traits” [28]. Several definitions emphasize the autonomous capabilities of these systems.

Of the 40 documents we analyzed, 15 documents (38%) did not provide any definition of AI at all, yet issued policy recommendations about AI. This finding affirms the need for a policy-facing definition of AI that would help adhere the growing number of guidance documents to technologies relevant to their implementation. In another case, policy guidelines from the AI Now Institute at New York University relied on Russell and Norvig’s [26] typology of definitional approaches by way of a definition, defining AI as “systems that think like humans, systems that act like humans, systems

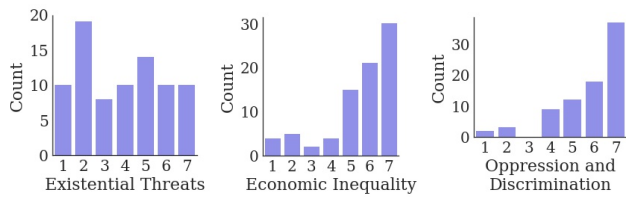


Figure 4: Histograms of AI researcher ratings of issues as relevant to AI. (1 = Strongly disagree, 7 = Strong agree)

that think rationally, systems that act rationally” [35]. Frequency of types encountered in our policy review appear in Figure 2. Example definitions sorted into the typology are shown in Table 2.

4.3 Quantitative Analysis

To avoid concerns about our qualitative coding procedure, we also replicated our analysis with a quantitative method using simple natural language processing. For this quantitative analysis, we use a keyword-based classifier to judge whether definitions are using humans as a comparison point for AI simply by searching for whether either definition provided contains or does not contain the word “human”. We find that 28% of the definitions given to us by AI researchers use the word “human”, while 62% of definitions from published policy documents include it.

4.4 Issue Analysis

Given that policy documents use more human-like definitions of AI, while AI researchers favor definitions that emphasize technical problem specification and functionality, which of these classes of definitions is more relevant to concerns people express about AI? Our first researcher survey included questions addressing this question. We analyzed what social issues AI researchers view to be relevant to AI; existential threats to humanity, economic inequality and the future of work, and oppression and discrimination (the first being commonly associated with human-like AI and the latter two with more mundane AI).

Our results show that there is a large degree of agreement that economic inequality (82% agree) and oppression and discrimination (83% agree) are relevant issues, both of which are commonly associated with existing technologies. There was more disagreement about whether existential threats are relevant (42% agree), which is an issue more relevant to (hypothetical) human-like AI. To further connect definitions of AI with issues about AI, we also asked a follow-up question to those participants who had rated any of the technologies we presented as AI and rated any of the issues we had presented as relevant: “Do any of the technologies you classified as A.I., or others that come to mind, relate to the issues you indicated were relevant to the topic of A.I.?” Figure 5 shows that substantial fractions of our sample considered Booking Photo Comparison Software and Siri to be relevant to oppression and discrimination, and a somewhat smaller fraction also considered those technologies to be relevant to economic inequality. Several AI researchers also rated Automated License Plate Readers as related to these two issues. These results show that many AI researchers

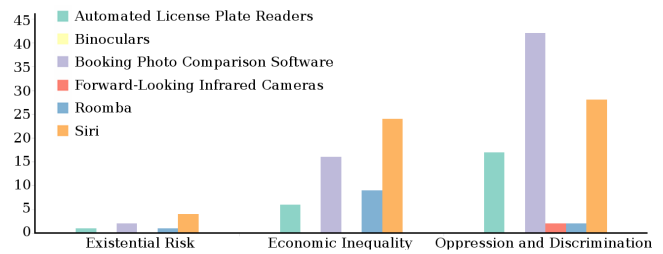


Figure 5: Frequency of survey responses indicating relevance of an AI technology to a particular issue.

consider existing, widely used information technologies to be AI and to be relevant to issues pertaining to the regulation of AI.

5 DISCUSSION

The motivation of this research was to describe how AI researchers and policymakers define AI and to examine possible gaps between their definitions. A disconnect between research and policymaker definitions may lead to harmful and unintended consequences in the realm of policymaking. Public concern about AI regarding issues such as economic inequality and existential risk call for policymakers to address these issues, but the absence of useful policy-facing definitions of AI hampers the development of appropriate and effective regulatory policies.

We found that relative to AI researchers, policy documents tend to favor human thinking/behavior in their definitions of AI. One possible consequence of this is that policy documents may overly focus on the future social consequences of AI. As noted by Cave et al., [8, 10], policymaker conceptions of humanoid AI may focus on the retreating horizon of systems still-to-be-created at the risk of passing over autonomous systems already in place.

To address this gap between policy-makers’ and AI researchers’ definitions of AI, and to practically address the social impacts of AI, we suggest using a definition that would maintain high fidelity to researcher definitions of AI while better lending itself to policy implementation. First, a policy-facing definition ought to fit with existing scholarly work on AI, and thus with AI researchers’ conceptions of AI. Second, this definition should reflect the concerns of citizens. Such a definition should capture, rather than leave out, AI technologies of worry to the general public. This would also mean pulling back the AI conversation from future-focused issues to focus on more immediate issues. Therefore, we propose the following necessary criteria for a policy-facing definition: (i) inclusivity of both currently deployed AI technologies and future applications, (ii) accessibility for non-expert audiences, and (iii) allowing for policy implementation of reporting and oversight procedures.

We conclude our analysis by offering an example of a recent policy definition that we believe meets these criteria. The OECD offers the following definition: “An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.”⁷ This definition meshes well with those in

⁷<https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449>

scholarly work such as given by Russell and Norvig [26] and more recently by Bryson [6] that characterize AI as systems aimed at accomplishing a particular goal by taking input from a digital or physical environment and undertaking goal-oriented action on that information by producing a decision or other output. In line with our criteria, the OECD definition is also specific enough to include existing technologies. For instance, a system like Automated License Plate Readers (ALPRs) makes recommendations that influence how police interact with their environment. Finally, also in line with our criteria, this definition emphasizes that the goals of AI are *human* goals, and not intrinsic to the AI itself (c.f. [20]).

6 CONCLUSION

Conversations about AI—including what it is and how it is affecting society—abound in policy discussions, technical research communities, and the public arena. Here, we highlighted differences between how AI researchers define AI technologies and how policy documents—including from organizations dedicated to the policy and ethics of AI—define AI technologies. We find that while AI researchers tend to define AI in terms of ideal thinking/behavior, policy documents tend to define AI in terms of human thinking/behavior. An important consideration in this comparison is that definitions adhering closely to the technical functionality of AI systems are more inclusive of technologies in use today, whereas definitions that emphasize human-like capabilities are most applicable to hypothetical future technologies. Therefore, regulatory efforts that emphasize AI as human-like may risk overemphasizing concern about future technologies at the expense of pressing issues with existing deployed technologies.

ACKNOWLEDGMENTS

Special thanks to many friends and colleagues for piloting, publicizing, and participating in our survey, and to Emma Spiro, Ceilyn Boyd, and danah boyd for helpful comments on our thoughts and methods. This work was supported in part by the Washington Research Foundation, and by a Data Science Environments project award from the Gordon and Betty Moore Foundation (Award #2013-10-29) and the Alfred P. Sloan Foundation (Award #3835) to the University of Washington eScience Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

REFERENCES

- [1] Philip E Agre. 1997. Lessons learned in trying to reform AI. *Social Science, Technical Systems, and Cooperative Work: Beyond the Great Divide* (1997), 131.
- [2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica*, May 23 (2016).
- [3] Peter Asaro. 2012. On banning autonomous weapon systems: Human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross* 94, 886 (2012), 687–709.
- [4] IEEE Standards Association et al. 2017. The IEEE global initiative on ethics of autonomous and intelligent systems. *IEEE.org* (2017).
- [5] Avrim Blum, Suriya Gunasekar, Thodoris Lykouris, and Nati Srebro. 2018. On preserving non-discrimination when combining expert advice. In *Advances in Neural Information Processing Systems (NeurIPS)*. 8376–8387.
- [6] Joanna J Bryson. 2019. The Past Decade and Future of AI's Impact on Society. *Towards a New Enlightenment: A Transcendent Decade*. Madrid: Turner-BVVA. (2019).
- [7] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*. 77–91.
- [8] Stephen Cave, Kate Coughlan, and Kanta Dihal. 2019. Scary robots: Examining public responses to AI. In *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*.
- [9] Stephen Cave, Claire Craig, Kanta Sarasvati Dihal, Sarah Dillon, Jessica Montgomery, Beth Singler, and Lindsay Taylor. 2018. *Portrayals and perceptions of AI and why they matter*. Technical Report. The Royal Society.
- [10] Stephen Cave and Seán Ó hÉigeartaigh. 2018. An AI race for strategic advantage: Rhetoric and risks. In *AAAI/ACM Conference on Artificial Intelligence, Ethics and Society*.
- [11] Ariel Conn. 2016. The problem of defining autonomous weapons. *The Future of Life Institute* (2016).
- [12] Information Technology Industry Council. 2017. *AI Policies and Principles*. (2017).
- [13] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *San Fransico, CA: Reuters*. Retrieved on October 9 (2018), 2018.
- [14] Paul Duckworth, Logan Graham, and Michael A Osborne. 2019. Inferring work task automatability from AI expert evidence. *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)* (2019).
- [15] Madeleine Clare Elish and danah boyd. 2018. Situating methods in the magic of Big Data and AI. *Communication Monographs* 85, 1 (2018), 57–80.
- [16] Ethan Fast and Eric Horvitz. 2017. Long-term trends in the public perception of artificial intelligence. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- [17] Luciano Floridi, Josh Cowsls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. 2018. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines* 28, 4 (2018), 689–707.
- [18] Anna Lauren Hoffmann. 2019. Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society* 22, 7 (2019), 900–915.
- [19] Bill Howe. 2019. Protect the public from bias in automated decision systems. *Seattle Times* (2019).
- [20] Deborah G Johnson and Mario Verdichio. 2017. Reframing AI discourse. *Minds and Machines* 27, 4 (2017), 575–590.
- [21] Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*. 4842–4852.
- [22] Donald Michie. 1971. *Formation and execution of plans by machine*. University of Edinburgh Department of Machine Intelligence & Perception.
- [23] Brent Mittelstadt. 2019. AI ethics—Too principled to fail? *arXiv preprint arXiv:1906.06668* (2019).
- [24] Vincent C Müller and Nick Bostrom. 2016. Future progress in artificial intelligence: A survey of expert opinion. In *Fundamental Issues of Artificial Intelligence*. Springer, 555–572.
- [25] Group of Governmental Experts. 2019. Group of Governmental Experts on Lethal Autonomous Weapons. *Fifth Review Conference of the High Contracting Parties to the Convention on Certain Conventional Weapons* (2019).
- [26] Stuart J Russell and Peter Norvig. 2016. *Artificial Intelligence: A Modern Approach*. Malaysia; Pearson Education Limited.
- [27] Science and Technology Committee. 2018. Algorithms in decision-making. *UK Parliament House of Commons* 36 (2018).
- [28] Personal Data Commission Singapore. 2019. A Proposed Model Artificial Intelligence Governance Framework. (2019).
- [29] Mario Luis Small. 2011. How to conduct a mixed methods study: Recent trends in a rapidly growing literature. *Annual Review of Sociology* 37 (2011), 57–86.
- [30] Brad Smith and Harry Shum. 2018. The future computed: Artificial intelligence and its role in society. *Microsoft Corporation* (2018).
- [31] Katta Spiel, Oliver L Haimson, and Danielle Lottridge. 2019. How to do better with gender on surveys: A guide for HCI researchers. *interactions* 26, 4 (2019), 62–65.
- [32] Latanya Sweeney. 2003. That's AI?: A history and critique of the field. (2003).
- [33] Moshe Y Vardi. 2012. Artificial intelligence: Past and future. *Commun. ACM* 55, 1 (2012), 5–5.
- [34] Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. *Discriminating Systems: Gender, Race, and Power in AI*. Technical Report. AI Now Institute.
- [35] Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kazianas, Mathur Varoon, Sarah Myers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. 2018. *AI Now 2018 Report*. Technical Report. AI Now Institute, New York, NY.
- [36] Meg Young, Michael Katell, and P. M. Krafft. 2019. Municipal surveillance regulation and algorithmic accountability. *Big Data & Society* (2019).