Public Value Principles for Secure and Trusted AI

Sukumar Ganapati
Florida International University, Miami FL, USA
ganapati@fiu.edu

Kevin C. Desouza
Queensland University of Technology, Brisbane, Australia
kevin.c.desouza@gmail.com

ABSTRACT

The objective of this paper is to establish the fundamental public value principles that should govern safe and trusted artificial intelligence (AI). Public value is a dynamic concept that encompasses several dimensions. AI itself has evolved quite rapidly in the last few years, especially with the swift escalation of Generative AI. Governments around the world are grappling with how to govern AI, just as technologists ring alarm bells about the future consequences of AI. Our paper extends the debate on AI governance that is focused on ethical values of beneficence to that of economic values of public good. Viewed as a public good, AI use is beyond the control of the creators. Towards this end, the paper examined AI policies in the United States and Europe. We postulate three principles from a public values perspective: (i) ensuring security and privacy of each individual (or entity); (ii) ensuring trust in AI systems is verifiable; and (iii) ensuring fair and balanced AI protocols, wherein the underlying components of data and algorithms are contestable and open to public debate.

CCS CONCEPTS

• **B7; Security and privacy** → Human and societal aspects of security and privacy; Social aspects of security and privacy.

KEYWORDS

Artificial Intelligence, Security, Public value, Public good

ACM Reference Format:
Sukumar Ganapati and Kevin C. Desouza. 2024. Public Value Principles for Secure and Trusted AI. In *25th Annual International Conference on Digital Government Research (DGO 2024), June 11–14, 2024, Taipei, Taiwan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3657054.3657086

1 INTRODUCTION

The adoption of artificial intelligence (AI) in general, and Generative AI in particular has grown at breakneck speed over the last decade. Although AI is not a new field, the latest developments in AI mark a significant departure from the beginnings of AI. The fundamental principles of machine learning algorithms have a long lineage. The newness lies in the corpus of unstructured big data that cognitive computing systems can leverage and the digital networked infrastructure that supports large-scale distributed computation. The newness also lies in AI's ability to perform tasks that are quite close to human capability, like distinguishing and recognizing images, being able to speak like humans, and recognize patterns in general. AI's capacity to mimic human beings, which has been a long-term perseverance of computer technologists, is close to realization in many use cases. Chatbots like Amazon's Alexa, Apple's Siri, are commonly given examples, but there are many more instances of AI use that extend beyond Chatbots. Figure 1 shows the evolution of AI with human like capabilities in recognizing and generating images, language, reading, and speech. These capacities have developed very quickly over the last decade.

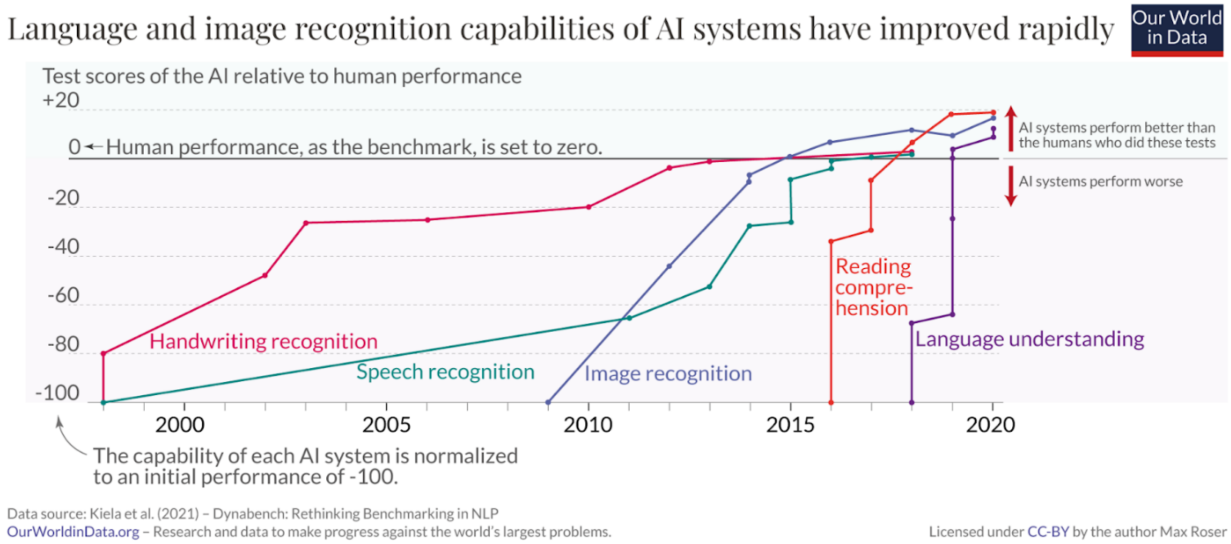
In parallel with the significant improvement in AI technology to mimic human beings, there is also breakneck growth in adopting these technologies across various sectors. ChatGPT, the most popular AI platform, reached 1 million users within 5 days in late 2022. Many large language models-based AI tools have emerged in parallel (e.g. Jasper Chat, Genie Chat, Dall-E, Bard, Bing AI, etc.). Venture capital investment in new AI startups has also boomed across different sectors in the last two years.

Given the quick growth of AI across different sectors, including government agencies, this article examines the policies for AI adoption. The main intent of the article is to identify the common public value principles that should ideally underlie AI adoption. While AI holds many economic benefits for the developers and startups investing in it, the larger public benefits and negative externalities of AI need to be carefully evaluated to identify the public values. AI should be accountable to citizen choices in this context [2]. AI systems can be deployed in a range of modes: fully autonomous, semi-autonomous, and in an augmented manner. The implementation of these AI systems within the context of public organizations needs to take into consideration the system's interaction with human beings: AI's use by decision-makers and AI's impact on the ultimate beneficiaries (i.e. the public). The public values approach is human centric in articulating such organizational use and impact of AI on human beings. It starts with the presumption of enhancing and maintaining mutually beneficial values, while minimizing harm.

The public values approach is a worthy model for the safe and secure adoption of AI. Public values, in this context, are normatively oriented toward preserving the rights of citizens, the obligations of AI in preserving such rights, and government policies which encourage such normative stance. If it is left to market forces, the current evolution of AI appears to focus on taking advantage of early commercial applications across different sectors. The private value of Open-AI and other entities (like Google and Microsoft) accrue to these creators only, even though they may purport to meet certain ethical dimensions in their implementation of AI. Considering AI as a public value provides insights into how the use and implications of AI are beyond the control of the creator. The public is left to face the many unintended consequences of AI implementation, so that the public bears more of the costs of these

DGO 2024, June 11–14, 2024, Taipei, Taiwan

ACM ISBN 979-8-4007-0988-3/24/06
https://doi.org/10.1145/3657054.3657086



Language and image recognition capabilities of AI systems have improved rapidly

The capability of each AI system is normalized to an initial performance of -100.

Data source: Kiela et al. (2021) – Dynabench: Rethinking Benchmarking in NLP
OurWorldinData.org – Research and data to make progress against the world's largest problems.
Licensed under CC-BY by the author Max Roser

Figure 1: Evolution of AI with human like capabilities [1]

downsides. AI is a public "beast" that is born and cannot be tamed by the creators themselves [3]. Even though the leading AI creators and over 1000 technologists called for a six-month moratorium on the development of AI in mid 2023, the AI genie was already out of the bottle. Taming the rapid evolution and use of GenAI for human benefit in different sectors requires a broader principled approach that is rooted in public values to humanize and democratize AI in the public domain context.

2 PUBLIC VALUES AND AI

The notion of public values is a contested one, with many flavors of what public value means [4]. Classical public administration literature has typically highlighted values like economy, efficiency, and effectiveness in designing and implementing programs; social equity has been added as a core dimension in recent years [5]. Transparency, accountability, and legitimacy are additional values that have persisted in the literature. Public value adds further dimensions to these core public administration values. Public values imply public interest and common good [6]. Public values are also associated with collectively desired social outcomes [7]. The public value framework is in contrast with new public management, which focuses on market-based solutions [8]. Instead, public value approach focuses on creating value for citizens. It has both traditional values (like efficiency and effectiveness) and emerging normative values (like justice, fairness, equality) [9]. Public value framework privileges the citizens' collective choices.

Nabatchi [10] presents four frames of public values for administration and governance: political, legal, organizational, and market. The political frame values democracy and public participation; this includes political representation, responsiveness, equality and liberty. The legal frame values individual rights, procedural due process, and equity. The organizational frame emphasizes bureaucratic values such as administrative efficiency, specialization and expertise, administrative structures and formal processes, organizational loyalty, and political neutrality. The market frame emphasizes economic values such as cost-savings, efficiency, productivity, entrepreneurship, innovation, and flexibility.

The public value framework has increasingly become a key concept for evaluating digital initiatives. Bannister and Connolly [11] provide a taxonomy of three public values by their orientation in the context of digital transformation: duty-oriented, service-oriented, and socially oriented. The duty-oriented values are the non-financial duties of a bureaucrat to the government and to the state; they include duties to citizens and elected officials, proper use of public funds, compliance with law, rectitude, integrity, and honesty. The service-oriented values are bureaucrat's responsibilities toward citizens incorporating services to such citizens in their various roles; delivering these services requires efficiency, effectiveness, responsiveness, transparency, and respect for the individual. The socially oriented values incorporate wider social goals of bureaucracy: they are inclusiveness, justice, fairness, impartiality, equality of treatment and access, due process, protection of citizen security, protection of citizen privacy, protection of citizens from exploitation, accountability, and consultation. Bannister and Connolly identify the potential impact that digital transformation can have on these public values.

Floridi and Cowls [12] identify an overarching framework consisting of five core principles for ethical AI: beneficence, non-maleficence, autonomy, justice, and explicability. The beneficence principle relates to promoting the well-being, while preserving human dignity and ecological sustainability. Non-maleficence is the principle of "do no harm" and "do only good"; it preserves the privacy and security of individuals. The autonomy principle relates to the power to decide, where the human autonomy overrides machine autonomy. The justice principle relates to promoting prosperity, preserving solidarity, and avoiding unfairness. The explicability principle includes the epistemological sense of intelligibility and transparency (as an answer to the question ‘how does it work?’) and in the ethical sense of accountability (as an answer to the question: ‘who is responsible for the way it works?’).

Ashok et al. [13] outline an initial framework for AI ethics, based on systematic literature review. They identify 14 digital ethics implications of AI use across four ontological frames: physical, cognitive, information, and governance. The physical domain has implications for new AI technology development, where the ethical principles are based on the beneficence principle of “do no harm”; these principles include dignity and well-being, safety, and sustainability. The cognitive domain is concerned with machine learning and algorithms, with ethical implications for fairness, promoting prosperity, and justice. AI, in this context, should be intelligible, accountable, explicable, and autonomous (i.e. AI should not hurt the power of humans to decide). The information domain is content related, where the key ethical principle is non-maleficence; it has ethical implications related to privacy and security of individuals. The governance domain is the regulatory context of rules and policies, with ethical implications.

Valle-Cruz & García-Contreras [14] posit that the main challenges of AI-driven public sector has to do with providing efficient and transparent services that provide public value and promote the benefit of society. AI could have a significant positive influence on public sector outcomes. The challenge for public agencies is to avoid the pernicious use of AI systems. AI-driven transformation and smart data management for the public sector are characterized by an operational change, which includes human resources and know-how as the spearhead. The transformation can be technological as long as strategic plans guarantee the ethical, transparent, and democratic use of emerging technologies.

Wirtz et al. [15] identify four major dimensions of challenges of implementing AI in the public sector context: AI technology implementation, AI law and regulation, AI ethics, and AI society. The major challenges for public agencies to implement AI technology include: safety of AI systems, the underlying system/data quality and integration, financial feasibility of implementing the AI systems, as well as specialization and expertise available within the government agencies to implement the systems. Governance challenges of AI are associated with legal and regulatory issues pertaining to data oversight and transparency, accountability and responsibility of algorithms, and privacy and safety of human beings (e.g. from security threats, surveillance). Public sector challenges with respect to AI ethics encompass consequences of AI rulemaking on human behavior, compatibility between machine and human value judgments, autonomous decision making when moral dilemmas exist, and AI discrimination (unfairness and inequality among people). The challenges of AI society include workforce substitution and transformation (with potential loss of jobs, social acceptance/trust in AI, and transformation of human-to-machine (H2M) and machine-to-machine (M2M) interaction.

The above literature shows the intense public debate surrounding the ethical use of AI, particularly in the public sector. The public sector has a special responsibility in the use of AI because of the very nature of government and bureaucracy. Governments cannot discriminate against individuals or groups in their provision of public services. Public agencies’ clientele and goods are non-excludable. No person can be legally refused from getting a service, unlike private services. Public goods are non-excludable and non-rival. Non-excludability implies that the government cannot prevent non-payers from consuming or using the good. Nonrival implies that one person’s consumption of the good does not prevent anyone else’s consumption of the good. Hence, there is free public access to the good.

Existing literature on AI has not theoretically considered the notion of AI as a public good that is non-excludable and nonrival. Much of the ethical debates of AI, as noted above, are framed from the moralistic approach of beneficence and harm. Taking this economic approach of public good, AI poses additional dimensions for public values. Gans [3] argues that socially minded technological entities will not be able to minimize AI’s harm from the unrestricted products released by for-profit firms. The development and use of AI could be considered as exemplary of market failure of a different nature from that of traditional public good. Whereas private market does not have incentives to create traditional public goods, the private tech startups did create AI drawing on open source and publicly available data. Unlike the tragedy of the commons where the common pool resources are depleted by the encroachment of private consumers, the nature of public data is such that the resource is not depleted despite its consumption by different AI actors. Rather, the data gets enriched and can be manipulated in creative ways, such that there are new synthetic data and recommendations. Such new synthesis holds both the upside (with predictions that hold broad public benefits, such as weather phenomena based on existing data) and the downside (with deep fakes and illusory data that are meant to intentionally mislead). AI thus does not face the tragedy of depletion of common pool data resources, but it faces the tragedy of plethora that can be put to unintended uses. While there are positive externalities of creating such large scale AI models, the negative externalities of the unintended uses of AI are beyond the control of the tech creators. Controlling AI use for public benefit would require a regulatory approach.

3 AI GOVERNANCE

In this section, we briefly consider how AI governance has been considered in two major rival contexts: the United States and Europe. AI implementation as well as policies are arguably more mature in these two democratic contexts. The United States has had a spate of AI technological development over the last few years; it is home to the major AI tech platforms like Open AI, Microsoft and Google. The European Union has taken a strong regulatory approach to technological developments in general over the last two decades, which has had influence beyond the EU. This “Brussels effect” is the global impact of EU tech policies; for example, the EU’s General Data Protection Regulation (GDPR) is the first such privacy legislation that has become a de facto standard globally [16]. The differences in the American and European approaches provides for interesting comparative empirical consideration of the policy contexts. It sheds light on the public value dimensions of AI policy and the extent to which the market failures of AI evolution are addressed. The empirical consideration combined with the theoretical consideration of AI as public good would be useful for providing fresh insights into the public value principles for governing AI more broadly.

3.1 United States

In the United States, there is no general legislative action yet from the federal government, even though there are several AI bills that have been promulgated. Rather, there are a few presidential initiatives on AI, undertaken as an executive action of the president's office. These executive actions largely focus on the federal government agencies' use of AI. Thus, the Executive Order 13960 of 2020 (Promoting the Use of Trustworthy AI in the Federal Government) squarely required federal agencies to inventory their AI use cases and share them publicly. The AI in Government Act of 2020 created an AI Center of Excellence within the General Services Administration and provided policies for AI acquisition and applications, including identifying best practices for mitigating AI risks.

The White House Office of Science and Technology Policy (OSTP) issued a Blueprint for an *AI Bill of Rights* in 2022, which laid out a set of five principles in the design, use, and deployment of AI systems. They are: safe and effective AI systems; algorithmic discrimination protections; data privacy; notice and explanation of AI use; providing human alternatives to opt out from machine decisions. The OSTP bill of rights touches on the public values principles of beneficence and not harming to some extent.

The Biden administration executive policies adopted in 2023 are largely supportive of private sector AI development. They do not necessarily prioritize the fundamental public value principles. Rather, the policies aim to blunt the harsh effects of AI on the edges. Thus, the AI Risk Management Framework (2023) undertaken by the National Institute of Standards and Technology is in collaboration with the private sector to better manage AI technological risks. The NIST framework is a voluntary guidance for the private tech sector to implement, aimed to improve trustworthiness considerations into AI design, development, use, and evaluation. The private sector tech firms are the ones guiding such guidelines. Subsequently, the Biden administration secured voluntary commitments from leading AI companies for safe, secure, and transparent development of AI technology. Lastly, the Executive Order 14110 (Safe, Secure, And Trustworthy Development and Use of Artificial Intelligence) of 2023 focused on AI safety and security. It included language on "responsible" innovation and competition, supporting American workers, preserving equity and civil rights and civil liberties, and providing consumer protection.

The American federal AI policies are primarily driven by the private sector AI interests. Public interests are included to the extent that they are in conformity with the private sector tech interests. These public interests are framed as "risk management" (e.g. NIST policies) and safe AI. Only the OSTP's bill of rights provides an explicit framework for public value principles. However, the bill of rights is not a formal policy measure; rather, it is an aspirational set of goals for AI implementation. So, there are early indications that the AI policies may not fully capture the public value principles. Or, at the least, they are shaped by the private interests of the tech sector over the public interests of AI use.

3.2 European Union

Unlike the United States, the European Union reached a political consensus on legislative action on the implementation and use of AI. Also, unlike the American policy framework that is shaped by the private tech sector, the EU framework promises to be broader based, taking the public values principles more seriously. The EU tech policies have generally focused on consumer benefits, rather than benefits for tech corporations. The Digital Markets Act (DMA) adopted in 2022 regulating "gatekeepers" is an exemplary policy in complementing EU competition policies by regulating large digital platforms (online search engines, app stores, messenger services). The Digital Services Act (enforced since 2024) regulates these platforms to prevent illegal and harmful activities online and the spread of disinformation. The DSA purports to rebalance the roles of users, platforms, and public authorities "according to European values, placing citizens at the centre."

The AI regulatory considerations in the EU have also been quite distinctive from the start, placing the citizens' rights ahead of machine capabilities. The High-Level Expert Group on AI's "Ethics Guidelines for Trustworthy Artificial Intelligence" (2019) laid out seven key requirements for trustworthy AI systems: empowering human agency and oversight; technical robustness and safety; privacy and data governance; transparency of data, system and AI business models; diversity, non-discrimination and fairness; social and environmental well-being; and accountability. The ethical requirements have since been reinforced in EU AI policy dialogues.

In 2021, the EU Commission put forward the "Proposal for a Regulation on AI" along with a "Coordinated Plan on AI." The proposed regulation explicitly acknowledged the "need to ensure a high level of protection of the public interests, in particular on health and safety, and people's fundamental rights and freedoms". The regulation focused on risks of AI's use, identifying four levels: unacceptable, high, limited, and minimal risks. The coordinated plan focused on AI investments, AI strategies and programs, and to align AI policies in order to remove fragmentation and address global challenges. These proposals provided a basis for the "horizontal AI regulatory framework" to harmonize rules for AI systems and their use across industries and sectors.

In 2023, the European Parliament and the Council reached a political agreement on the Artificial Intelligence Act, based on the risk based regulatory proposal. Unacceptably risky AI systems that pose a threat to fundamental rights (e.g., certain predictive policing applications, social scoring) are banned. High-risk applications (e.g. biometrics) will be required to comply with strict requirements of risk-mitigation, high quality data sets, human oversight, and a high degree of robustness. Minimal risk applications (e.g. AI-enabled recommendations) do not have regulatory obligations. AI generated content (e.g., synthetic audio, video, text and images) which could mislead people (e.g. deep fakes) need to be transparent and identified as such. Agencies that do not meet the regulatory requirements would be fined.

Clearly, EU AI policies have followed a different trajectory from that of the United States. The EU policy approach is fundamentally human centric, taking a risk-based approach to how AI impacts people. The AI policies have an ethical underpinning, providing for the individual privacy and safety. There is a horizontal approach to

AI risks, prohibiting high level risky AI uses while facilitating low risk applications.

4 PUBLIC VALUE PRINCIPLES FOR AI

The US and EU approach to AI policies provide insights into the underlying base from which the AI policies are set in both contexts. The US AI approach arguably begins from a private sector tech-based market approach, with the policies slanted toward and shaped by the tech sector. The EU AI approach begins from a human ethics approach, providing for a risk-based assessment of AI use. These two sets of policies have different sets of commitments toward public values. Arguably, the US approach is less aligned with public values and more aligned to market values. The EU approach is more aligned with public values. Yet, both do have elements of public values: they are both based on the values of beneficence and non-malfeasance.

Despite their differences, we argue that neither the US nor the EU policies are oriented to the new reality of AI as the public good that is nonexcludable. In this vein, the US market-based approach is inadequate in providing guidance on how to deal with the unintended negative externalities of AI. Indeed, as the past few months of large-scale language models and other uses show, there are severe negative externalities (e.g. deep fakes) which have emerged. The EU ethics-based approach is inadequate in providing guidance on how to build AI systems that can be secure and trustworthy. Simply banning high risk AI applications (like deep fakes) and allowing low risk uses do not lead to the development of more trusted AI. We need robust mechanisms to ensure the development of secure and trustworthy AI.

We extend on the upsides of the US and EU approaches to add a further set of principled mechanisms which could arguably provide a pathway for overcoming their downsides in a robust AI system. These principles stem from the very basic concepts of security and trust embedded in public values approach. First, security would entail safeguarding the subject (individual or organization), free from any unintended consequences of AI use. It entails the maintenance of individual autonomy and privacy. Second, trustworthiness implies the high degree of integrity of the underlying AI data and algorithms. It entails that trust can be independently verified rather than taken for granted. Third, protocols for AI applications (i.e. creating new content by machines) should be fair and balanced. They should benefit all the parties mutually, resulting in overall public benefits. We draw on game theory to show how cooperation could occur for optimal public benefits.

4.1 Ensuring security of subject

Ensuring the security and privacy of each individual (or entity) is crucial in the globally connected online world, where almost two-third of the world's population has access to the Internet [17]. The global outreach holds not only positive, but also negative externalities of AI as a public good. On the upside, the cost of information flow has flattened. On the downside, even if a person intends to be offline, there could be a footprint of the person online because of an online acquaintance. A citizen-centric approach to AI essentially entails preserving the security and privacy of the subject. Security implies that the individual is not harmed by the generative or predictive AI systems. The harm may not only be due to manipulation of legitimate information by unscrupulous actors, but also due to inaccurate information that gets reinforced. Privacy refers to the preservation of the individual's sensitive information beyond the reach of those who legitimately need the access. A person's privacy could be compromised in the online world not only because of the individual's actions, but also due to actions of those who possess the individual's information (e.g. data brokers, service providers, etc. who need access to legitimate information).

The public good approach to AI puts a different light on security and privacy from that of an ethical/ moral approach. The public good approach is an instrumental approach, which puts a utilitarian public value. The moral approach is a rights based approach, where the value is non-negotiable. The instrumental perspective is ultimately consequential in its evaluation, whereas the moral approach is deontological in its evaluation. The instrumental approach gives control to the subject for oversight on the individual's preference for the extent to which the information is revealed. The control should be negotiable for the individual to the extent that s/he benefits from sharing the information. That is, the individual does not only control, but also obtains a portion of the benefit of sharing the information. Neither the U.S. or the E.U. approach offers such access to benefit–they only provide less (e.g. in the U.S.) or more (in the E.U) control. Their approach is pecuniary in as much as they tax the AI platform private revenues. They do not have a theoretical underpinning for division of public value among individuals. An instrumental approach would democratize the benefits rather than merely providing access to control. Although platforms may still dominate the data access and distribution, the instrumental approach distributes the broad public value of AI among individuals.

As the data for AI can be potentially distributed across different sectors, realizing the public benefits requires a whole of government approach that can span across the public, nonprofit, and private sectors. With a market approach, the data for AI can be monetized by the entities holding the data. Unfortunately, individuals do not have similar ability to arrogate some of the benefits of the information they share. With the public benefits approach, the individuals would also be considered as equivalent to the entities holding the valuable personal data that they are sharing in exchange for a share in the benefits. A major value proposition of these systems is to tailor and personalize experiences and services that are commensurate with the benefits. The value could theoretically range from transparency of the data use to that of sharing material benefits. For example, in the case of Estonia, where the government has access to individual information and can query it, citizens are informed whenever someone uses their data. There is also an emerging argument of the AI Dividend, whereby AI platforms pay a dividend to individuals for their data use [18].

4.2 Ensuring mutual trust

Trust is a multidimensional concept [19]. At its very basic level, trust takes a long time to develop, but can be broken very quickly. From a game theory perspective, repeated games between the same players can lead to mutual understanding and consequent trust or distrust of each other. However, in the open digital world, the games are not repeated between individuals. Rather, the games are repeated between individuals and the digital platforms (e.g. in social media). Then, AI based digital platforms have the responsibility to establish trust. Trust can be broken down into three elements—ability, integrity, and benevolence. These elements need careful consideration when contemplating the public value of AI [20]. First, one must trust that the AI platform has the ability to perform effectively (i.e. correctly). In cases, where there is a high-degree of belief in the system's ability, it might be deployed in an autonomous manner with requisite human supervision. Second, integrity of AI systems ensures that they behave in a manner that is aligned with the social ethical expectations. In societies where one has the right to contest decisions and seek recourse, their interaction with cognitive computing systems should not limit or restrict these options. Benevolence is the belief that the cognitive computing systems have good intentions and are looking to advance the interest of each citizen they interact with while uploading the rule of law and adhering to administrative protocols.

The process of ensuring trust in AI platforms needs deeper consideration. Reputation of AI platforms can be a mechanism. It provides a means for indirect trust in two player settings of repeated or one-shot games. AI platforms face social scrutiny and gain or lose trust. This indirect trust reinforces reputation toward either side. Well established platforms that have broad appeal and have gone through multiple testing develop the reputation over time. But, trust can also be lost quickly with a few intended or unintended "mistakes" of the platforms. Competition between the platforms will then weed out the less reputable platforms (or platforms with bad reputations).

The above two player game assumes that there are no spillover effects of the game over other parties beyond that of reputation. Spillover effects that are harmful to other parties not in the game (e.g. generative AI that materially impacts third party outside of the players and the platform) cannot be controlled only by reputation. These will require the intervention of a third-party arbiter to monitor negative externalities. Triangulation offers a methodological path to ensuring trust in such conditions. Similar to Burt's [21] argument about building social capital, brokerage and closure can reinforce trust. Brokerage is the third-party connection between actors in a network who can fill the structural holes with their vision. Closure is the tightening of coordination in a closed network of people. Brokerage and closure can mediate the players to establish the trust. Third party reputation checks—by either within the tech sector or by public agency outside of the tech sector—can act as the mechanisms for brokerage and closure.

4.3 Ensuring fair and balanced protocols

AI fundamentally depends on two components, which determines the fairness and balance of its protocols: the underlying data on which the AI is trained, and the algorithm that the AI uses for making predictions. If any of these data or algorithms themselves are biased, the AI algorithms themselves get biased results. However, in many cases, detecting the bias may not be technically feasible. The datasets, even if they are large, may not represent the population. This will naturally lead to outcomes that can be prejudicial and unfair to certain groups (e.g. LGBTQ or First Nations). Hence, to ensure that these models generate value it is important to ensure that these systems work with humans who come from under-represented groups. This improves AI systems in two ways: first, to check the output, and second, to provide feedback to the systems to improve their performance.

To ensure fair and balanced protocols, there are two central requirements. First, the data and the algorithms should be transparent with mechanisms to critique their application. That is, the protocols should be contestable by the public; they cannot be black-box, proprietary mechanisms where the results are opaque to the public. Contestability allows humans to intervene and interrogate critical elements of AI protocols from conception to deployment. Similar to the numerous open debates that happen before public policy decisions, the public should be able to participate and provide input into the datasets used and be able to question algorithmic decisions since they affect individual lives in unintended ways [20].

Second, the issue of explainability is critical. Here, it is important to distinguish teleologically on what one needs an explanation for. Explainability is not the same as transparency, i.e. it is not only a catalog of what the underlying data and algorithms are. Rather, the explainability ideally refers to: (1) who has been involved in making the decision (human, machine, or both); (2) why the choice of decision-maker is in keeping with public value (i.e., there are economies of scale, humans-are-over-the-loop to supervise, etc); (3) explaining the process that the analytical machine used—here one can describe the inputs that went in and the overall analytical reasoning; and (4) the output, i.e. the predicted or generated content that may underlie a public decision (and who one can go to contest the decisions).

5 CONCLUSION

Our intent with this paper is not to provide any definitive one-shot policy strategies in the evolution of AI. Indeed, any attempt to provide such strategies would be futile in nature since the field is rapidly evolving. Along with the rapid evolution of AI, the means to tame the AI beast are also getting intensified. As the literature on debates about AI policies show, the public value approach does have a place in shaping the AI policies. The public value approach has mainly taken an intrinsic rights approach that assumes the foundational nature of AI, focusing on ethical and moral values of data use. We take a more instrumental approach in this paper to argue that AI should be viewed as a public good that is non-excludable. The public goods approach has the advantage of providing fresh insights into AI policies.

Leading AI policy entrepreneurs like the U.S. and the E.U. have taken different approaches in this context. The U.S. AI policies are mainly in response to the private tech sector and are shaped by the tech entrepreneurs. They are aligned with the private benefits for the tech platforms; public values are espoused mainly on the edges to blunt the harsh effects of the tech platforms' approach. The E.U. policies are more citizen and human centric, taking a rights-based approach of who owns the data. They are better aligned to traditional public values, imposing horizontal regulations on the tech platforms.

We argue that neither the U.S. nor E.U. have taken an instrumental approach. The instrumental approach could provide a more nuanced insight into the AI policies. We need robust mechanisms to ensure the development of secure and trustworthy AI. For this, we have advanced three arguments. First, security and privacy of each individual (or entity) is crucial. The sharing of information should hold dividends for those individuals. That is, AI's public value should be democratized. Second, trust in AI systems should be enforceable and verifiable; the trust cannot be taken for granted. Third, the AI protocols need to be fair and balanced, wherein the underlying components of data and algorithms are contestable and open to public debate. The components should not only be transparent, but also explainable.

ACKNOWLEDGMENTS

Sukumar Ganapati acknowledges funding support from National Science Foundation Grant 1924154 (CyberTraining: Implementation: Medium: Advanced Cyber Infrastructure Training in Policy Informatics)

REFERENCES

[1] Max Roser. (2022, December 6). *The Brief History of Artificial Intelligence: The World has Changed Fast – What Might be Next?* Our World in Data. Retrieved February 6, 2024, from https://ourworldindata.org/brief-history-of-ai

[2] Gerry Stoker. (2006). Public value management: A new narrative for networked governance? *The American Review of Public Administration*, *36*(1), 41-57. https://doi.org/10.1177/0275074005282583

[3] Joshua S. Gans. (2023). Can Socially-Minded Governance Control the AGI Beast? *NBER Working Paper No. 31924.* JEL No. L20,O33,O36. Retrieved February 6, 2024, from https://www.nber.org/system/files/working_papers/w31924/w31924.pdf

[4] Prudence R. Brown, Lorraine Cherney, and Sarah Warner. (2021) Understanding Public Value – Why Does It Matter?, *International Journal of Public Administration*, 44 (10), 803-807, https://doi.org/10.1080/01900692.2021.1929558

[5] Kristen Norman-Major. (2011). Balancing the Four Es; or Can We Achieve Equity for Social Equity in Public Administration? *Journal of Public Affairs Education*, 17(2), 233–252. https://doi.org/10.1080/15236803.2011.12001640

[6] Barry Bozeman. (2007). *Public Values and Public Interest: Counterbalancing Economic Individualism.* Georgetown University Press, Washington DC.

[7] Mark H. Moore. (1995). *Creating Public Value: Strategic Management in Government.* Harvard University Press, Cambridge, MA.

[8] Antonio Cordella and Carla M. Bonina. (2012). A Public Value Perspective for ICT Enabled Public Sector Reforms: A Theoretical Reflection. *Government Information Quarterly*, 29(4), 512-520. https://doi.org/10.1016/j.giq.2012.03.004.

[9] John M. Bryson, Barbara C. Crosby, and Laura Bloomberg. (2014), Public Value Governance: Moving Beyond Traditional Public Administration and the New Public Management. *Public Administration Review*, 74 (4), 445-456. https://doi.org/10.1111/puar.12238

[10] Tina Nabatchi. (2018). Public Values Frames in Administration and Governance, *Perspectives on Public Management and Governance*, 1 (1), 59–72. https://doi.org/10.1093/ppmgov/gvx009

[11] Frank Bannister and Regina Connolly. (2014). ICT, Public Values and Transformative Government: A Framework and Programme for Research. *Government Information Quarterly*, 31 (1), pp. 119-128. https://doi.org/10.1016/j.giq.2013.06.002.

[12] Luciano Floridi and Josh Cowls. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, 1(1). https://doi.org/10.1162/99608f92.8cd550d1

[13] Mona Ashok, Rohit Madan, Anton Joha, and Uthayasankar Sivarajah. (2022). Ethical Framework for Artificial Intelligence and Digital technologies. *International Journal of Information Management*, 62, 102433. https://doi.org/10.1016/j.ijinfomgt.2021.102433

[14] David Valle-Cruz and Rigoberto García-Contreras. (2023). Towards AI-Driven Transformation and Smart Data Management: Emerging Technological Change in the Public Sector Value Chain. *Public Policy and Administration*, 0(0). https://doi.org/10.1177/09520767231188401

[15] Bernd W. Wirtz, Jan C. Weyerer and Carolin Geyer. (2019) Artificial Intelligence and the Public Sector—Applications and Challenges, *International Journal of Public Administration*, 42(7), 596-615. https://doi.org/10.1080/01900692.2018.1498103

[16] Anu Bradford. (2020). *The Brussels effect: How the European Union Rules the World.* Oxford University Press, New York.

[17] International Telecommunication Union (ITU). (2023). *The Global Connectivity Report 2022.* Retrieved February 6, 2024, https://www.itu.int/itu-d/reports/statistics/global-connectivity-report-2022/

[18] Barath Raghavan and Bruce Schneier. (2023, 29 June). Artificial Intelligence Can't Work Without Our Data. *Politico*, Retrieved February 6, 2024, from https://www.politico.com/news/magazine/2023/06/29/ai-pay-americans-data-00103648

[19] Roger C. Mayer, James H. Davis, David F.. Schoorman. (1995). An Integrative Model of Organizational Trust, *Academy of Management Review*, *20* (3), 709-734. https://doi.org/10.2307/258792

[20] Kevin C. Desouza and Gregory S. Dawson. (2023). *Pathways to Trusted Progress with Artificial Intelligence.* IBM Center for the Business of Government. https://www.businessofgovernment.org/sites/default/files/Pathways to Trusted Progress with AI.pdf.

[21] Ronald S. Burt. (2005). *Brokerage and Closure: An Introduction to Social Capital.* Oxford University Press, New. York.