Subject: Book Proposal Development Grant Submission Confirmation
Date: Monday, August 30, 2021 at 7:09:40 PM Eastern Daylight Time
From: Princeton University Press
To: Baobao Zhang

Hello and thank you for your book proposal development grant submission.

Your application submission has been received by the Princeton University Press grant committee. The application cycle will close August 30, 2021. Applicants will be notified of decisions the week of October 18, 2021.

Submitted on Mon, 08/30/2021 - 19:08
Submitted by: Anonymous
Submitted values are:
First Name
Baobao

Last Name
Zhang

Preferred Pronouns
She/Her/Hers

Email
bzhang22@syr.edu

Affiliation
Maxwell School of Citizenship and Public Affairs, Syracuse University

Author C.V./Resume
zhang-baobao.pdf

Proposed project title
AI Governance in Our Age of Distrust

Project description
Overview:

As artificial intelligence (AI) has become more widely deployed in workplaces, schools, hospitals, and homes, many are concerned about the realized and potential harms of the technology. The phrase “trustworthy AI” has become ubiquitous in AI ethics statements published by governments, tech companies, and civil society groups. Although the definition of trustworthy AI varies, common principles include beneficence, non-maleficence, autonomy, justice, and explicability [1]. The buzz around trustworthy AI has produced a trove of research on designing algorithms that the public will trust. Proposals typically call for technical solutions such as developing models to align with human values, auditing models for bias and safety, and creating documentation of training data and models [2].

While these technical solutions are important, focusing on them exclusively ignores human users' subjective perceptions and experiences of those impacted by AI systems. This book project argues that building and deploying “trustworthy AI” requires the public to trust the institutions behind these AI systems. Unfortunately, distrust in institutions, including governments and tech companies, has become widespread around the world. I take a mixed-methods approach using surveys, experiments, and case studies to show how institutional distrust could hinder both the regulation of AI systems and the use of beneficial AI.

Audience:

I plan to write this book to reach an audience both in academia and beyond. The book will interest scholars and students from disciplines as diverse as public policy, political science, information science, science and technology studies, sociology, law, and computer science. Instructors can assign the book for courses on technology policy/law, AI ethics, computer science and ethics, and human-computer interaction. Given the timeliness of the subject matter, this book will also attract a readership among policymakers, journalists, and civil society groups.

Table of contents:

Chapter 1: This introductory chapter explains how current proposals for building trustworthy AI from tech companies and governments focus extensively on technological standards while neglecting institutional trust.

Chapter 2: This chapter theorizes that the public's trust in AI cannot be divorced from their trust in tech companies, governments, and other actors building and deploying the technology. Furthermore, I discuss the concept of reputational spillover: the general reputation of an institution can influence how the public view AI built or deployed by that institution.

Chapter 3: This chapter provides empirical evidence through original survey and experimental data (from the US and the EU), along with case studies, to show that trust in AI systems is linked to trust in the institutions building and deploying the technology. My research findings also raise a dilemma for the governance of AI: the public sometimes places more trust in tech companies than the governments that are tasked with regulating the technology [3].

Chapter 4: Ability, benevolence, and integrity are three factors theorized to increase perceived trustworthiness [4]. Through case studies, I show that tech companies, governments, journalists, and civil society groups are contesting the assessment of these three qualities in specific AI systems and AI in general. As a result, the public becomes uncertain whether they can trust specific AI systems or AI in general.

Chapter 5: This chapter explores how public distrust -- sometimes warranted and sometimes not -- in governments, scientists, and technology has led to poor outcomes in climate change mitigation and the Covid-19 response. I draw lessons from these two policy challenges and apply them to AI governance.

Chapter 6: This chapter highlights the two potential dangers of deploying AI systems in a low-trust environment: coercion and cynicism. Although the public may not trust unsafe or discriminatory AI systems, they are compelled to be surveilled and evaluated by these systems through economic or political coercion. On the other hand, cynicism could lead the public to feel unmotivated to fight for their rights or to reject AI applications that could be beneficial to society. I conclude by proposing solutions that could empower the public to demand a say in AI governance.

References:

[1] Thiebes, Scott, Sebastian Lins, and Ali Sunyaev. "Trustworthy artificial intelligence." Electronic Markets (2020): 1-18.

[2] Brundage, Miles, et al. "Toward trustworthy AI development: mechanisms for supporting verifiable claims." arXiv preprint arXiv:2004.07213 (2020).

[3] Zhang, Baobao, and Allan Dafoe. "US public opinion on the governance of artificial intelligence." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020.

[4] Mayer, Roger C., James H. Davis, and F. David Schoorman. "An integrative model of organizational trust." Academy of Management Review 20.3 (1995): 709-734.

Subject area

Politics and International Relations

Book Coach Partners
Jane Joann Jones, Ph.D., Margy Thomas, Ph.D.