Public Opinion Toward Artificial Intelligence

Baobao Zhang

May 2021

Abstract

This chapter synthesizes and discusses research on public opinion toward artificial intelligence (AI). Understanding citizens' and consumers' attitudes toward AI is important from a normative standpoint because the public is a major stakeholder in shaping the future of the technology and should have a voice in policy discussions. Furthermore, the research could help us anticipate future political and consumer behavior. Survey data worldwide show that the public is increasingly aware of AI; however, they – unlike AI researchers – tend to anthropomorphize AI. Demographic differences correlate with trust in AI in the abstract: those living in East Asia have higher levels of trust in AI, while women and those of lower socioeconomic status across different regions have lower levels of trust. Surveys that focus on particular AI applications, including facial recognition technology, personalization algorithms, lethal autonomous weapons, and workplace automation, add complexity to this research topic. I conclude this chapter by recommending three new directions for future studies: understanding 1) how institutional reputation affects trust in AI, 2) how increasing one's experience with and knowledge about AI affects attitudes, and 3) how attitudes toward AI shapes individuals' behavior.

Keywords: artificial intelligence, public opinion, survey research

This essay will appear as a chapter in the *Oxford Handbook of AI Governance*, which I am co-editing.

1 Introduction

Public opinion toward artificial intelligence (AI) has become an emerging area of study within AI policy. Much of the existing survey research has been conducted by companies, think tanks, and governments rather than academics. Nevertheless, the growth of AI ethics as a field of study has increased the number of academic publications on the topic. Furthermore, survey work has expanded beyond high-income countries to include respondents in the Global

South. This chapter synthesizes and discusses research on public opinion toward AI, at the same time proposing new directions for research.

Understanding public opinion toward AI matters for AI governance for two reasons. First, from a normative perspective, the public is a major stakeholder in shaping the future of AI and, therefore, should be included in discussions around AI governance. Secondly, in cases of other technologies (e.g., nuclear energy and genetically modified foods), the public could shape the development and deployment of AI. Understanding what they think about AI will help us anticipate future political contestation and consumer behavior.

Even as AI systems become more widely deployed in public, most of the existing work in AI ethics focuses on ethics principles developed by tech companies, governments, think tanks, or academic institutions (Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020). Much of the work is prescriptive: they describe what ideal "ethical" AI systems should look like. These principles, such as respecting human rights or preserving privacy, aim to protect the public's welfare. Nevertheless, public input to shape these ethical principles has been limited or costly. Participatory design has been suggested as one way to ensure that AI systems work in the public's interest (Kulynych et al., 2020). While public opinion research may not directly impact computer scientists' design choices, it can help inform policymakers and tech companies of the public's concerns about AI in general or specific applications of AI. For instance, the research could illuminate concerns that are not salient in elite discourses about AI governance or reveal consensus around an ethics principle.

Furthermore, survey research could illuminate how the public has conflicting views about what they consider ethical uses of AI. Indeed, some of the published AI ethics principles conflict with each other (Whittlestone, Nyrup, Alexandrova, & Cave, 2019). For instance, the lack of non-white people in image databases used to train computer vision systems produces less accurate predictions for darker-skinned individuals. Nevertheless, gathering additional training data could mean greater surveillance of populations that are overly policed in high-income countries. These theoretical tensions also play out in disagreements between survey

respondents. In the US, support for facial recognition technology varies significantly by race, party identification, and age group (Smith, 2019). The Moral Machine project, which collected 40 million decisions in 233 countries and territories, reveals that respondents in different regions and cultures have different preferences regarding autonomous vehicles' behavior in moral dilemmas (Awad et al., 2018). This type of research necessarily complicates AI ethics by revealing that consumers and voters disagree about how AI systems should be developed and deployed.

The public has shaped technology policies, including genetically modified (GM) foods, nuclear power, and vaccines. Understanding public sentiments could help policymakers, activists, and tech companies anticipate mass mobilization around AI-related issues, particularly around calls to ban specific AI applications. First, they act as direct consumers and boycott products or services produced by the technology. For instance, European consumers perceive GM foods as risky and have been reluctant to consume them (Frewer et al., 2004). In the past two decades, vaccine confidence has declined in many parts of the world (de Figueiredo, Simas, Karafillakis, Paterson, & Larson, 2020), leading to stagnating vaccination rates (Requejo, Griffiths, Duncan, Mirza, & Mebrahtu, 2020). Second, the public can demand change in regulation through mass mobilization, activism, and voting. For instance, anti-vaccine groups in the US have pushed states to adopt nonmedical exemptions for vaccines (Olive, Hotez, Damania, & Nolan, 2018). Voters opposing nuclear power have voted in referendums to ban (Austria 1978, Italy 2011) or phase out (Switzerland 2017) nuclear power (Pelinka, 1983; Moody, 2011; BBC, 2017).

Those opposed to what they perceive to be unethical AI-adjacent technologies have adopted similar tactics to limit or ban the use of these technologies. For instance, organizations like the Electronic Frontier Foundation have discouraged consumers from purchasing Amazon's Ring home security system to protect user privacy and prevent excessive police surveillance (Guariglia, 2020). Furthermore, members of the public have increasingly protested against AI-adjacent applications deployed by the state. In 2020, hundreds of stu-

dents in the UK protested outside the Department for Education, decrying an algorithm that predicted their exam grades and was unfair to students from lower socioeconomic backgrounds (Kolkman, 2020). As a result, officials reversed course and threw out the grades predicted by the algorithm. As AI and AI-adjacent applications become more widely deployed, consumer and citizen mobilization may become more widespread.

Here, I lay out the structure of this chapter. First, I summarize survey findings regarding the public's knowledge of AI and general trust in AI. I break down these results by country and by demographic subgroups, including gender and socioeconomic status. Secondly, I consider public attitudes toward four specific applications of AI: facial recognition technology, personalization algorithms, lethal autonomous weapons, and workplace automation. These four applications currently have high political salience around the globe; as a result, there exists a large trove of survey data on these topics. Finally, I conclude this chapter by discussing three topics for further research: institutional trust in actors behind AI systems, the impact of experience and knowledge on attitudes toward AI, and how beliefs about AI impact consumer and civic behavior.

2 Fundamental public opinion research: knowledge and trust

2.1 Knowledge about AI

Two of the most fundamental questions in studying public attitudes toward AI include how much the public knows about the technology or how they define it. While the public may not have complete technical knowledge of AI, a 2018 survey conducted in eight low-, middle-, and high-income countries suggests that the vast majority of respondents have heard of AI (Kelley et al., 2019). Furthermore, this and other studies have demonstrated that the public has at least partial knowledge of what AI is by describing machines making decisions typically

made by humans or mentioning related technologies like robots (Cave, Coughlan, & Dihal, 2019). While computer scientists tend to define AI by its technical functionality, the public's definition of AI tends to compare it with human behavior or intelligence. Furthermore, the public perceives AI as futuristic rather than something that they already interact with. These trends may be driven by how the news media and popular media depict AI systems. Recent criticism of existent algorithms and AI applications may change the public's understanding of the technology.

One standard textbook definition of AI is "the study of agents that receive precepts from the environment and perform actions," where "each such agent implements a function that maps precept sequences to actions" (Russell & Norvig, 2020). A 2019 survey of AI and machine learning (ML) researchers found that 72% of respondents preferred definitions of AI that emphasized mathematical problem solving and technical functionality over definitions that compared machines with humans (Krafft, Young, Katell, Huang, & Bugingo, 2020). Although computer scientists prefer to define AI without emphasizing comparisons with humans, popular understanding of the technology tends to anthropomorphize AI (Salles, Evers, & Farisco, 2020). For instance, in a 2018 nationally representative survey in the UK, 42% of respondents "referred to computers performing tasks that replicated aspects of human cognition" and 25% referred to robots when describing AI (Cave et al., 2019). In a 2018 nationally representative survey in the US, respondents were more likely to label tech applications that can socially interact with humans (e.g., virtual assistants, social robots) as AI than applications that cannot (e.g., Google Translate, Google Search). (Zhang & Dafoe, 2019).

Why the public anthropomorphizes AI could be explained by the media they consume. Across eight countries, the three top ways that the public learns about AI are through social media, TV reports and commentaries, and movies or TV shows (Kelley et al., 2019). A majority of English-language policy documents published by governments, tech companies, and civil society groups defined AI in comparison to human cognition or behavior (Krafft

et al., 2020). A review of AI in fictional narratives finds that AI is frequently depicted as intelligent machines embodied in humanoid forms (Cave et al., 2018). AI ethicists worry that representing AI as human-like could lead the public to be misinformed about the risks and benefits of AI (Cave et al., 2018). One particular concern is that the public fails to recognize how AI systems are deployed today, most of which are not humanoid robots but commonplace software (Krafft et al., 2020).

Related to the issue of anthropomorphizing AI, the public appears to be subject to the "AI effect." The AI effect is the phenomenon in which people perceive an AI system not to be "truly intelligent" once it solves a problem; as a result, AI is viewed as a futuristic technology rather than an existent one (McCorduck & Cfe, 2004). Content analysis of open-ended responses from the public in eight countries revealed that 24% described AI as "futuristic," frequently referring to science fiction (Kelley et al., 2019). In the previously discussed 2018 survey of the US public, a majority of respondents assume that Facebook photo tagging, Google Search, Netflix or Amazon recommendations, and Google Translate do not use AI (Zhang & Dafoe, 2019). In contrast, the majority of AI/ML researchers surveyed in (Krafft et al., 2020) consider automated license plate readers and booking photo comparison software to use AI. The AI effect may eventually fade as existing applications of the technology, such as facial recognition software and large language models (e.g., OpenAI's GPT-3), become more salient in the news. Furthermore, criticism of predictive analytics that is not as advanced as AI, such as the UK grading algorithm and software used to make welfare decisions (Eubanks, 2018), are now incorporated in discussions around AI ethics and governance.

2.2 Trust in AI systems

The phrase "trustworthy AI" has become an ubiquitous phrase 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 (Thiebes, Lins, & Sunyaev, 2020). The buzz around trustworthy AI has produced various theoretical literature on designing algorithms and institutions that the public will trust. The chief shortcoming of these works is that they de-emphasize human users' subjective perceptions and experiences of those impacted by AI systems. At the same time, public opinion research has not kept up with these theoretical contributions; instead, they focus on the public's general attitude toward AI divorced of any technical or institutional context. Bridging the gap between these related areas of study would enhance our understanding of trustworthy AI.

Theoretical works on trustworthy AI frequently lay out socio-technical frameworks for building AI systems that the public will trust. These frameworks propose solutions that often involve building technical or institutional features to promote safety, fairness, and transparency. Some proposed solutions include offering explanations for how the AI system works, creating documentation of the development process, requiring third-party audits, offering rewards for those who identify flaws in AI systems, creating a database of AI incidents, and enacting regulation (Brundage et al., 2020; Jacovi, Marasović, Miller, & Goldberg, 2021; Knowles & Richards, 2021). While these solutions could theoretically make AI systems safer and more ethical, their emphasis on the technical overlooks the role of subjective human judgment. In the classic "integrative model of organizational trust" (Mayer, Davis, & Schoorman, 1995), ability, benevolence, and integrity are factors that increase *perceived* trustworthiness (emphasis mine). Furthermore, adding to the notion that trust is subjective, the authors argue that "[p]eople differ in their inherent propensity to trust" (715). Consider the example of explaining how AI systems work: empirical studies find that people with different levels of expertise have different levels of comprehension of the same explanations (Saha et al., 2020). Furthermore, even subject area experts were persuaded to trust AI systems when given misleading explanations (Lakkaraju & Bastani, 2020). Finally, individual differences in personality or technical expertise are correlated with different baselines level of trust in AI systems (Oksanen, Savela, Latikka, & Koivula, 2020).

In contrast, much of the public opinion research has focused on respondents' subjective evaluation of AI in general. While these survey questions ask about respondents' perceived impact of AI on society or their support for developing AI, they could be measuring the general level of trust in AI. One central flaw in these studies is that they often ask about AI as a technology devoid of context, such as how or where the AI system is deployed. Nevertheless, they highlight differences in trust by country, gender, and socioeconomic status.

Those living in East Asia, compared with other regions, have greater trust in AI across several comparative surveys. In a 2019 survey of over 150,000 respondents in 142 countries, 59% of those in East Asia indicated that AI would mostly help society while 11% indicated that AI will mostly harm society. In contrast, in Latin America and the Caribbean, the region most wary of AI, 49% indicated that the technology will mostly help society while 26% indicated that it will mostly harm society (Neudert, Knuutila, & Howard, 2020). These results have been replicated in another cross-national survey showing that those living in East Asian countries view the development of AI and workplace automation most positively (Johnson & Tyson, 2020). Content analysis of open-ended responses found that 14% of responses from South Korea describe AI as "worrying," compared with 30% in the US and 31% in France (Kelley et al., 2019). In the US and the EU, where trust in AI systems are mixed, there is widespread consensus that AI is a technology that should be carefully managed (Zhang & Dafoe, 2020; Eurobarometer, 2017).

Two other important trends observed in these studies are that women and those of lower socioeconomic status (e.g., lower levels of education, lower income) are less trusting of AI. In 15 out of 20 countries surveyed in (Johnson & Tyson, 2020), men, compared with women, have significantly more positive view of AI development. In the same study, those with higher levels of education (having completed post-secondary education in high-income countries; having completed secondary education in middle-income countries) have more positive view of AI development. In the US, those who earn more than $100,000 annually, compared with those in lower income brackets, have the highest level of support for developing AI at 59%.

In contrast, only 33% of those earning less than $30,000 annually support developing AI (Zhang & Dafoe, 2019). Across 142 countries, business executives and other white-collared professionals, compared with those engaged in manual labor, are more likely to perceive AI as being helpful to society (Neudert et al., 2020). More research is needed to investigate how these differences in attitudes toward AI formed. Nevertheless, research has identified ways that AI systems have disproportionately harmed women, non-white people, and those of low socioeconomic status and excluded them from deciding how the technology is built and deployed (Gebru, 2020).

Given the novelty of AI as a topic for public opinion research, most cross-national studies ask about general attitudes toward AI devoid of technical or institutional context. Nevertheless, these studies reveal important variations in trust toward AI by country, gender, and socioeconomic status. Future survey work could examine how various subgroups in different countries perceive proposed strategies to make AI systems more trustworthy (e.g, by offering explanations or performing third-party audits). An additional advance in research would consider what drives perceived trust in AI systems deployed in different settings (e.g., facial recognition versus tagging abusive online content). While AI is often called a general-purpose technology, most AI systems deployed today have narrow applications. In the next section, I explore public opinion research that does not focus on AI in general but uses three specific applications: facial recognition, lethal autonomous weapons, and workplace automation.

3 Views toward four applications of AI

This section discusses the public opinion research on four highly salient applications in the news and has generated greater interest among survey researchers. These examples illustrate the need to study public opinion toward specific applications of AI, rather than AI as a general concept, as respondents often rely on their existing heuristics when considering applications of AI in daily life.

3.1 Facial recognition

Facial recognition algorithms used to identify, verify, and classify persons based on their facial features have been deployed in at least 98 countries (Gosh, 2020). The technology has been standard in consumer applications, such as unlocking smartphones or tagging people in photos; now, law enforcement, employers, and businesses are increasingly turning to the technology as well. As facial recognition becomes more widespread, civil society groups and academic researchers have pointed out flaws in these AI systems and the risk to privacy and civil liberties. Researchers found that leading commercial facial recognition software programs are much less accurate at identifying women, particularly those with darker skin, than white men (Buolamwini & Gebru, 2018). Even if facial recognition algorithms were to become more accurate, critics contend that the technology would increase the capacity of law enforcement, governments, and even private companies to monitor the public — causing disproportionate harm to already marginalized groups (West, Whittaker, & Crawford, 2018).

The public's view toward facial recognition technology is nuanced, although some key trends have replicated across surveys. First, in several countries, the public is more supportive of facial recognition technology used by law enforcement compared with private actors. Second, support is correlated with demographic variables, such as the resident country, race, and political leaning.

Although much of the criticism of facial recognition technology has focused on its use and abuse by law enforcement, adults in several countries are more supportive of its use by law enforcement than businesses or employers. In the US, 59% of adults find uses by law enforcement to assess security threats in public spaces to be acceptable, while only 15% find uses by advertisers to track responses to ads acceptable (Smith, 2019). In the UK, 70% of adults support uses in criminal investigations and in airports to verify travelers; in contrast, only 7% support uses by supermarkets to track shopper behavior, and 4% support uses by employers to evaluate job candidates (Ada Lovelace Institute, 2019). Similarly, in Australia, more than 70% of adults support uses by the police. In contrast, less than a

quarter of Australian adults support uses by businesses to track customers or advertise to them (Automated Society Working Group, 2020). In a 2019 study of the public in China, Germany, the UK, and the US, support for central governments' use of facial recognition technology is higher than support for use by private businesses (Kostka, Steinacker, & Meckel, 2021).

Demographic variables, including country of residence, race, and political leaning, correlate with support for facial recognition technology. In the four-country survey discussed above, respondents in China indicated the highest support for facial recognition use (67% support). In contrast, 38% of adults in Germany, 50% in the UK, and 47% in the US support the use of the technology (Kostka et al., 2021). Respondents in China, compared with the other countries, perceive facial recognition technology to be more convenient and efficient as well as less risky from privacy, discrimination, and surveillance considerations. This finding is perhaps not surprising given the prevalence of the technology deployed by law enforcement and businesses in China.

Beyond these variations between countries, there are also differences in support among demographic subgroups within a country. For example, while more than a majority of US adults support law enforcement's use of facial recognition, support is much lower among Black Americans and those who identify with or lean toward the Democratic Party (Smith, 2019). One explanation is that these demographic subgroups also have lower trust in law enforcement in general. Furthermore, US cities and states that have banned or placed a moratorium on the police's use of facial recognition technology are left-leaning in their politics (Recognition, 2020). As criticism of the technology becomes more politically salient, public opinion could change. For example, in the US, opposition to the use of facial recognition software has increased by 16 percentage points between 2018 and 2019 (Sabin, 2019).

3.2 Personalization algorithms

Personalization algorithms employ a user's online or offline data to create online content tailored to the user. Applications that use personalization algorithms include recommendation systems for news articles on social media, targeted online advertising, or individualized pricing. While personalization algorithms have become ubiquitous, researchers, journalists, and civil society groups have pointed out harms from the technology. From a privacy perspective, users' personal and behavioral data are collected without their informed consent by tech companies to sell to advertisers (Zuboff, 2019). Targeted advertising has excluded women and ethnic minorities from job listings and rental listings (Imana, Korolova, & Heidemann, 2021; Spinks, 2019). While concerns about "filter bubbles" on social media are overblown for the majority of users, personalization could be reinforcing the views of those with extremist political beliefs (Stark, Stegmann, Magin, & Jürgens, n.d.). This subsection reviews research that examines how users themselves understand and view personalization algorithms. Although public opinion research on this topic is growing, qualitative research has provided additional insights that address puzzles posed by survey data.

The works reviewed highlight the vast information asymmetry between the public and the tech companies that build and deploy personalization algorithms. The public lacks knowledge about the technology or even lacks the vocabulary to talk about how these algorithms affect their online experiences. For example, nearly three-quarters of US Facebook users do not know that Facebook assigns them "interest categories" that are used to recommend them ads, news, and other content (Hitlin & Rainie, 2019). Furthermore, the majority of the US public do not perceive Facebook automated photo tagging or Netflix/Amazon's recommendation systems to involve AI or ML (Zhang & Dafoe, 2019). Qualitative interviews with 22 young people aged 16 to 26 revealed these frequent users of social media do not have a clear understanding of how personalization algorithms work (Swart, 2021). The researcher notes that while these young users lack the technical vocabulary to talk about personalization algorithms, the algorithms are objectively non-transparent. Qualitative interviews with

"power users" (those who use a privacy/security browser extension to track and block data collection) revealed that even those with high levels of technical knowledge do not think they fully understand the personalization algorithms they are resisting (Kant, 2020).

Although the public may not fully understand how personalization algorithms work, they oppose certain types of data from being collected and used or certain types of personalization. Data from nationally representative surveys in Germany, Great Britain, and the US find that the publics in these countries oppose tech companies collecting sensitive information like personal tragedies or household incomes; in addition, there is a consensus against using personalization in political campaigning (Kozyreva, Lorenz-Spreen, Hertwig, Lewandowsky, & Herzog, 2021). These survey results replicate findings from a survey of 748 Amazon Mechanical Turk workers that revealed respondents perceive the use of household income and race, compared with other types of data, in algorithmic personalization to be the most unfair (Coen, Paul, Vanegas, Lange, & Hans, 2016).

The US public is not as opposed to the personalization of online newspaper front pages as those in Germany and Great Britain (Kozyreva et al., 2021). Nevertheless, 62% of US adults said that social media companies have too much control over the news people see, and 55% said these companies create a worse mix of news (Shearer & Grieco, 2019). Breaking down the data by party identification, Republicans express more negative views about social media platforms, with 85% indicating that these platforms censor their viewpoints — compared with 62% of Republicans (Smith, 2018). In reality, researchers do not find empirical evidence that platforms like Facebook, Twitter, or YouTube censor conservative viewpoints; in fact, these platforms go out of their way to appease conservatives in the US (Barrett & Sims, 2021).

Another gap between perception and reality is that many users want some personalization; yet, at the same time, they oppose tech companies collecting their data that are needed to build the personalization algorithms. Seven-four percent of the public in Germany, along 62% of the public in Great Britain and the US, exhibit this "acceptability gap" between

personalized online services and data collection (Kozyreva et al., 2021). The researchers acknowledge that the public may not be aware that building personalization algorithms require collecting user data. Another view posited by Kant (2020) suggests that users are negotiating difficult trade-offs between protecting their privacy and accessing convenient online services, at the same time, acknowledging that personalization algorithms are highly opaque and data collection is impossible to circumvent.

3.3 Lethal autonomous weapons

Lethal autonomous weapon systems can identify and engage targets without human intervention. The use of an autonomous drone (made by a Turkish company) in a March 2020 skirmish in Libya could possibly be the first deployment of this technology in battle (Vincent, 2021). Civil society groups, including the Campaign to Stop Killer Robots, have advocated for an international ban on fully autonomous weapons. These groups argue that lethal autonomous weapons are unethical and unsafe; furthermore, they suggest that an arms race to develop the technology would exacerbate tensions between major military powers. At the same time, the low cost of building lethal autonomous weapons could lead to proliferation among non-state actors, including terrorists (Warren & Hillas, 2020). Thirty countries have publicly expressed support for a pre-emptive international ban on fully autonomous lethal weapons. Still, several major military powers, including the US and Russia, currently oppose such a ban (Campaign to Stop Killer Robots, 2019).

Human Rights Watch and the Campaign to Stop Killer Robots have conducted two cross-national surveys that examine attitudes toward lethal autonomous weapons. While these organizations take an explicit policy position regarding the technology, their survey samples, compared with those in academic studies, contain the most diverse respondents in terms of geography. In 2018, 61% of those surveyed in 26 countries opposed the use of lethal autonomous weapons while 22% support their use (Deeney, 2019). In a similar study conducted in 2019, 56% surveyed expressed opposition while 24% expressed support.

Considerable cross-national variations exist: the 2018 survey found that support for fully autonomous weapons is highest in India (50%) and Israel (41%) and lowest in Turkey (13%) and Hungary (13%).

Academic studies find that US adults' support for lethal autonomous weapons can be affected by framing or new information. A 2013 survey experiment found that those who consume science fiction oppose lethal autonomous weapons when they are primed to think about films that feature killer robots (Young & Carpenter, 2018). Two survey experiments conducted in 2015 find that informing the US public that lethal autonomous weapons would be used to protect US troops increased support for developing the technology (Horowitz, 2016). The same paper reveals that informing the US public that foreign countries or non-state actors are developing these weapons also increased support. Future research could consider how other types of messaging would affect attitudes toward lethal autonomous weapons or expand the respondent pool to consider non-US publics.

3.4 Workplace automation

Concerns about workplace automation have existed throughout the 20th century but have recently intensified with the increasing focus on AI. According to the OCED, 14 percent of the jobs in 32 OECD countries are at high risk of being automated in the coming decades (Nedelkoska & Quintini, 2018). A more dire forecast puts the number at 47% of US jobs (Frey & Osborne, 2017). Survey data over time show that the US public's views toward workplace automation have become more uncertain in recent years. The US National Science Foundation (NSF) had conducted eight surveys between 1983 and 2003, asking respondents whether they agree or disagree that computers and factory automation will create more jobs than they will eliminate. Survey data from (Zhang & Dafoe, 2019) showed similar levels of disagreement with the NSF survey statement (around half of the respondents disagreed) but a higher percentage of respondents who indicated they do not know (24% in 2018 versus less than 10% in all of the NSF surveys).

Recent survey research has attempted to disentangle the fear of automation in general versus the fear of one's own job becoming automated. Workers in the US express optimism bias regarding automation: they believe that while many jobs are likely to be automated, their own will be safe from automation (Smith & Anderson, 2016). Respondents whose jobs are objectively more likely to be automated do not think that their jobs are at higher risk of automation. The correlation between workers' forecasts and some economic forecasts is as low as 0.11 (Zhang, 2019). Furthermore, workers indicate they are more worried about losing their jobs to cheaper labor than being replaced by computers and machines (Smith, 2016).

One set of findings among recent observational studies is that actual or anticipated exposure to automation is positively correlated with support for right-wing populist parties, candidates, or policies. Comparative analysis using regional-level and individual-level data from several European countries find that those more exposed to automation shocks indicated greater support for nationalist and radical-right parties (Anelli, Colantone, & Stanig, 2019; Im, Mayer, Palier, & Rovny, 2019), even after accounting for social welfare programs that potentially helped workers harmed by automation (Gingrich, 2019). In the US, exposure to industrial robots is positively correlated with support for Donald Trump in the 2016 Presidential Election at the electoral district level (Frey, Berger, & Chen, 2018). Individual-level survey data suggests that Americans who are more exposed to automation express greater opposition to free trade and immigration (Wu, 2019).

The link between fear of automation and right-wing politics is not replicated across all studies. Other studies find that workers exposed to automation risks are more in favor of left-wing policies or parties – yet another study finds that the threat of automation does not shift political preferences at all. An analysis of survey data from 17 European countries between 2002 and 2012 finds that respondents whose jobs were more automatable expressed greater support for redistribution (Thewissen & Rueda, 2019). Exposure to automation is positively correlated with support for not only radical right-wing parties but also mainstream left-wing

parties (Gingrich, 2019). In a survey experiment, an informational treatment that made American respondents more aware of automation's threat *increased* support of universal basic income (UBI) among low-skilled workers (Lekalake, Markovich, Nahmias, & Russell, 2019). Other studies find null effects. In another survey experiment, exposing US workers to news articles about how automation will threaten jobs in general and their individual jobs did not shift support for expanding the welfare state or UBI (Zhang, 2019). In an observational study of 21 European countries, researchers found no association between risk of job automation and UBI support (Dermont & Weisstanner, 2020).

Given workers' uncertainty about how AI will impact their jobs, it seems reasonable that this set of nascent research papers would reach different conclusions. A shortcoming of these research papers is that they focus on OECD countries while ignoring workers in the Global South. Future survey research should consider expanding the geographic scope by surveying workers in middle and low-income countries.

4 Directions for future research

This chapter attempts to systemically review the existing research on the public's attitudes toward AI. An increasing number of cross-national studies allow researchers to explore variations in attitudes between countries and demographic subgroups. Researchers have also branched out to study specific applications, such as facial recognition technology, lethal autonomous weapons, and workplace automation. Nevertheless, there is much potential for future research that expands upon existing studies. Here I propose three new directions.

First, researchers could explore institutional trust in AI within the contemporary political and economic context. This line of research somewhat differs from empirically testing the theories of trustworthy AI, which tend to generate abstract solutions and de-emphasize the power struggle between tech companies, governments, and the public. Given current policy debates around regulating major tech companies and technological competition between the

US and China, empirical studies must not be agnostic to the actors in this space. In fact, survey research in the US shows that the public has different levels of trust in actors to build AI systems (Zhang & Dafoe, 2019). US adults place the greatest amount of trust in university researchers and the military to build AI; furthermore, they place greater trust in tech companies than the government. However, trust in tech companies is not uniform: the public places significantly less trust in Facebook than other major tech companies.

Future research could try to explain such variations in trust in actors building AI systems. Recently, there has been increasing backlash against major tech companies over their disproportionate market dominance as well as their failure to protect user data and prevent the spread of dis/misinformation. Research questions could examine whether the overall reputation of a tech company affects the public's perception of its AI products. For instance, would the public place less trust in AI systems developed by a company that has repeatedly reported data breaches or has poor content moderation practices?

The second research direction is to examine how increasing users' experience with and knowledge about AI will impact their attitudes toward the technology. Many governments and civil society groups have proposed educating the public about AI to empower citizens. Furthermore, various national AI strategies have called for educating students about AI to train a competent workforce for the future. Finally, as more AI systems become deployed in the real world, the public will increasingly interact with AI applications online, in public, or at their workplaces. This review has shown that making generalizations about how increased experience and knowledge impact attitudes toward AI is difficult. Further theoretical and empirical work should take a more nuanced approach.

For instance, the relationship between technical knowledge about AI and trust in AI systems does not appear to be monotonic increasing. Those without technical training tend to be more distrustful, while those with some technical training (e.g., those who have computer science or engineering degrees) tend to be more trusting (Zhang & Dafoe, 2019). Yet, AI/ML researchers appear to be increasingly aware of the dangers and negative societal

consequences of AI systems (Zhang et al., 2021; Belfield, 2020). Future research could test whether knowledge about AI and trust in AI systems follows an inverted-U shape: increasing knowledge increases trust up to a point then decreases as one becomes an expert.

Another aspect of this research direction is to examine how different types of experience or education impact attitudes toward AI. The impact of experience on trust varies by the type of AI the user interacted with, according to a systemic literature review (Glikson & Woolley, 2020). For virtual AI and embedded AI (AI that is embedded in socio-technical systems and not visually visible to users), trust starts high and decreases with use. In contrast, for robotic AI, trust starts low and increases with use. The authors of the review piece acknowledge these trends are typically observed in short-term studies and propose that future research track how experience impact trust with long-term use. Research has also shown that using an AI application does not necessarily increase knowledge of how AI systems work. For example, those who frequently use social media do not understand how platforms use their data to generate personalized content to categorize users (Hitlin & Rainie, 2019; Swart, 2021). Therefore trust might be mediated through subjective user experience and not necessarily a technical assessment of the AI's capability or safety.

Finally, when educating people about AI, the type of information conveyed (e.g., technical knowledge versus information about the societal impact of AI) could affect attitudes differently. Papers presented at ML conferences tend to focus on improving the performance, generalization, and efficiencies of models (Birhane et al., 2021), rather than making AI systems safer, fairer, or more explainable – topics that are prevalent at AI ethics conferences (Spinks, 2019). Future research could test how taking a course on AI ethics or attending an AI ethics conference (versus taking a computer science class on AI or attending an ML conference) would impact students' trust in AI and their views toward AI ethics.

The third research direction for future research is to consider how beliefs about AI impact consumer and civic behavior. More and more AI systems are embedded within products or services that the public can purchase or deployed in public spaces. Researchers studying

human trust in AI have used short-term, small-sample experiments to test whether explaining how the AI makes decisions will increase trust (Glikson & Woolley, 2020). Deploying these experiments in real-world settings and taking multiple measures over time could help researchers understand how stated preferences relate to behavior. For instance, would informing consumers about the risks and benefits of using an AI-powered mental health chatbot affect people's willingness to use it? Would educating people about racial/gender bias in facial recognition technology increase the likelihood of their signing a petition to ban its use by law enforcement?

This third line of inquiry could draw upon human-computer interaction research on privacy and consumer behavior. One central finding in this literature is the privacy paradox: although consumers indicate that they care about privacy, they take little effort to protect their privacy online (Barth & De Jong, 2017). More recent work pushes back against the idea that consumers are irrational by arguing that privacy policy statements are too complex for consumers to understand (Bashir, Hayes, Lambert, & Kesan, 2015) and that consumers have grown too cynical about websites and apps' willingness or ability to protect their data (Hoffmann, Lutz, & Ranzini, 2016). We could observe a similar "AI ethics paradox" in future studies where the public expresses deep concerns about AI systems causing harm yet fail to take action as consumers or citizens. If anything, the information asymmetry between developers and the public is even more significant in the AI realm than in the privacy realm: black-box algorithms are far more incomprehensible to laypeople than lengthy privacy policies.

The three new research directions discussed above are just some ways that scholars of public opinion can advance research on this topic. Researchers can draw upon the ever expanding literature on AI ethics and governance from ML, information science, and science and technology studies to develop new surveys and experiments.

Acknowledgements

I am grateful for feedback from Laurin Weissinger, Toby Shevlane, and Nataliya Nedzhvetskaya. I would also like to thank Aura Gonzalez for her helpful research assistance.

References

Ada Lovelace Institute. (2019). *Beyond face value: public attitudes to facial recognition technology* (Tech. Rep.). Ada Lovelace Institute. Retrieved from `https://perma.cc/M4EG-N5MY`

Anelli, M., Colantone, I., & Stanig, P. (2019). *We were the robots: Automation in manufacturing and voting behavior in western europe.* Retrieved from `https://www.iza.org/publications/dp/12485/we-were-the-robots-automation-and-voting-behavior-in-western-europe` (Working Paper)

Automated Society Working Group. (2020). *Australian attitudes to facial recognition: A national survey* (Tech. Rep.). Monash University. Retrieved from `https://www.monash.edu/__data/assets/pdf_file/0011/2211599/Facial-Recognition-Whitepaper-Monash,-ASWG.pdf`

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64.

Barrett, P. M., & Sims, J. G. (2021). *False accusation: The unfounded claim that social media companies censor conservatives* (Tech. Rep.). New York University, Stern School of Business, Center for Business and Human Rights. Retrieved from `https://perma.cc/X5A8-HQLB`

Barth, S., & De Jong, M. D. (2017). The privacy paradox–investigating discrepancies between expressed privacy concerns and actual online behavior–a systematic literature review. *Telematics and informatics*, *34*(7), 1038–1058.

Bashir, M., Hayes, C., Lambert, A. D., & Kesan, J. P. (2015). Online privacy and informed consent: The dilemma of information asymmetry. *Proceedings of the Association for Information Science and Technology*, *52*(1), 1–10.

BBC. (2017). *Switzerland votes to phase out nuclear power.* Author. Retrieved from `https://www.bbc.com/news/world-europe-39994599` (Accessed: 10 May 2021)

Belfield, H. (2020). Activism by the ai community: Analysing recent achievements and future prospects. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 15–21).

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2021). The values encoded in machine learning research. *arXiv preprint arXiv:2106.15590*.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., ... others (2020). Toward trustworthy AI development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91).

Campaign to Stop Killer Robots. (2019). *Country views on killer robots.* Author. Retrieved from https://www.stopkillerrobots.org/wp-content/uploads/2019/10/KRC_CountryViews_25Oct2019rev.pdf (Accessed: 10 Jun 2021)

Cave, S., Coughlan, K., & Dihal, K. (2019). Scary robots: Examining public responses to AI. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 331–337).

Cave, S., Craig, C., Dihal, K., Dillon, S., Montgomery, J., Singler, B., & Taylor, L. (2018). *Portrayals and perceptions of ai and why they matter* (Tech. Rep.). The Royal Society. Retrieved from https://royalsociety.org/-/media/policy/projects/ai-narratives/AI-narratives-workshop-findings.pdf

Coen, R., Paul, E., Vanegas, P., Lange, A., & Hans, G. (2016). *A user-centered perspective on algorithmic personalization* (Tech. Rep.). Retrieved from https://perma.cc/AE2Q-PRWB

Deeney, C. (2019). *Six in ten (61%) respondents across 26 countries oppose the use of lethal autonomous weapons systems.* (Available at https://www.ipsos.com/en-us/news-polls/human-rights-watch-six-in-ten-oppose-autonomous-weapons [accessed: 14 Jun. 2021])

de Figueiredo, A., Simas, C., Karafillakis, E., Paterson, P., & Larson, H. J. (2020). Mapping global trends in vaccine confidence and investigating barriers to vaccine uptake: a large-scale retrospective temporal modelling study. *The Lancet*, *396*(10255), 898–908.

Dermont, C., & Weisstanner, D. (2020). Automation and the future of the welfare state: basic income as a response to technological change? *Political Research Exchange*, *2*(1), 1757387.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor.* New York: St. Martin's Press.

Eurobarometer. (2017). *Special Eurobarometer 460: Attitudes towards the impact of digitisation and automation on daily life* (Tech. Rep.). Author. Retrieved from https://perma.cc/9FRT-ADST

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI* (Tech. Rep.). Berkman Klein Center for Internet & Society. Retrieved from https://ssrn.com/abstract=3518482orhttp://dx.doi.org/10.2139/ssrn.3518482

Frewer, L., Lassen, J., Kettlitz, B., Scholderer, J., Beekman, V., & Berdal, K. G. (2004). Societal aspects of genetically modified foods. *Food and Chemical toxicology*, *42*(7), 1181–1193.

Frey, C. B., Berger, T., & Chen, C. (2018). Political machinery: Did robots swing the 2016 us presidential election? *Oxford Review of Economic Policy*, *34*, 418–442.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, *114*, 254–280.

Gebru, T. (2020). Race and gender. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of AI Ethics* (pp. 253–269). Oxford, UK: Oxford University Press.

Gingrich, J. (2019). Did state responses to automation matter for voters? *Research & Politics*, *6*(1), 2053168019832745.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660.

Gosh, I. (2020). *Mapped: The state of facial recognition around the world* (Tech. Rep.). Visual Capitalist. Retrieved from `https://www.visualcapitalist.com/facial-recognition-world-map/`

Guariglia, M. (2020). *What to know before you buy or install your Amazon Ring Camera.* Electronic Frontier Foundation. Retrieved from `https://www.eff.org/deeplinks/2020/02/what-know-you-buy-or-install-your-amazon-ring-camera` (Accessed: 10 May 2021)

Hitlin, P., & Rainie, L. (2019). *Facebook algorithms and personal data* (Tech. Rep.). Pew Research Center. Retrieved from `https://perma.cc/S7YG-PPA5`

Hoffmann, C. P., Lutz, C., & Ranzini, G. (2016). Privacy cynicism: A new approach to the privacy paradox. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *10*(4).

Horowitz, M. C. (2016). Public opinion and the politics of the killer robots debate. *Research & Politics*, *3*(1). Retrieved from `https://doi.org/10.1177/2053168015627183`

Im, Z. J., Mayer, N., Palier, B., & Rovny, J. (2019). The "losers of automation": A reservoir of votes for the radical right? *Research & Politics*, *6*(1), 2053168018822395.

Imana, B., Korolova, A., & Heidemann, J. (2021). Auditing for discrimination in algorithms delivering job ads. In *Proceedings of the Web Conference 2021* (pp. 3767–3778).

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 624–635).

Johnson, C., & Tyson, A. (2020, 12). *People globally offer mixed views of the impact of artificial intelligence, job automation on society* (Tech. Rep.). Pew Research Center. Retrieved from `https://www.pewresearch.org/fact-tank/2020/12/15/people-globally-offer-mixed-views-of-the-impact-of-artificial-intelligence-job-automation-on-society/`

Kant, T. (2020). *Making it personal: Algorithmic personalization, identity, and everyday life.* Oxford, UK: Oxford University Press.

Kelley, P. G., Yang, Y., Heldreth, C., Moessner, C., Sedley, A., Kramm, A., ... Woodruff, A. (2019). "Happy and assured that life will be easy 10years from now.": Perceptions of artificial intelligence in 8 countries. *arXiv preprint arXiv:2001.00081*.

Knowles, B., & Richards, J. T. (2021). The sanction of authority: Promoting public trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 262–271).

Kolkman, D. (2020). *F**k the algorithm?: what the world can learn from the uk's a-level grading fiasco.* London School of Economics and Political Science. Retrieved from `https://blogs.lse.ac.uk/impactofsocialsciences/2020/08/26/fk-the-algorithm-what-the-world-can-learn-from-the-uks-a-level-grading-fiasco/` (Accessed: 10 May 2021)

Kostka, G., Steinacker, L., & Meckel, M. (2021). Between security and convenience: Facial recognition technology in the eyes of citizens in China, Germany, the United Kingdom, and the United States. *Public Understanding of Science*, 09636625211001555.

Kozyreva, A., Lorenz-Spreen, P., Hertwig, R., Lewandowsky, S., & Herzog, S. M. (2021).

Public attitudes towards algorithmic personalization and use of personal data online: evidence from germany, great britain, and the united states. *Humanities and Social Sciences Communications*, *8*(1), 1–11.

Krafft, P. M., Young, M., Katell, M., Huang, K., & Bugingo, G. (2020). Defining AI in policy versus practice. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (p. 72–78). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3375627.3375835 doi: 10.1145/3375627.3375835

Kulynych, B., Madras, D., Milli, S., Raji, I. D., Zhou, A., & Zemel, R. (2020). *Participatory approaches to machine learning.* International Conference on Machine Learning Workshop.

Lakkaraju, H., & Bastani, O. (2020). "How do i fool you?" manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 79–85).

Lekalake, R., Markovich, Z., Nahmias, G., & Russell, S. (2019). *Automation risk and support for a universal basic income.* (Working Paper)

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734.

McCorduck, P., & Cfe, C. (2004). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence.* CRC Press.

Moody, B. (2011). *Italians say no to nuclear energy in referendum.* Reuters. Retrieved from https://www.reuters.com/article/uk-italy-nuclear/italians-say-no-to-nuclear-energy-in-referendum-idUKTRE75C3P020110613 (Accessed: 10 May 2021)

Nedelkoska, L., & Quintini, G. (2018). *Automation, skills use and training* (Tech. Rep.). OECD iLibrary. Retrieved from https://doi.org/10.1787/2e2f4eea-en

Neudert, L.-M., Knuutila, A., & Howard, P. (2020). *Global attitudes towards AI, machine learning & automated decision making* (Tech. Rep.). Oxford Internet Institute. Retrieved from https://oxcaigg.oii.ox.ac.uk/wp-content/uploads/sites/124/2020/10/GlobalAttitudesTowardsAIMachineLearning2020.pdf

Oksanen, A., Savela, N., Latikka, R., & Koivula, A. (2020). Trust toward robots and artificial intelligence: An experimental approach to human–technology interactions online. *Frontiers in Psychology*, *11*.

Olive, J. K., Hotez, P. J., Damania, A., & Nolan, M. S. (2018). The state of the antivaccine movement in the united states: A focused examination of nonmedical exemptions in states and counties. *PLoS medicine*, *15*(6), e1002578.

Pelinka, A. (1983). The nuclear power referendum in Austria. *Electoral Studies*, *2*(3), 253–261.

Recognition, B. F. (2020). *Ban facial recognition map.* (Available at: https://www.banfacialrecognition.com/map/ [accessed: 1 Jun. 2021)

Requejo, J., Griffiths, U., Duncan, R., Mirza, I., & Mebrahtu, S. (2020). *Immunization coverage: Are we losing ground?* (Tech. Rep.). UNICEF. Retrieved from https://data.unicef.org/resources/immunization-coverage-are-we-losing-ground/

Russell, S., & Norvig, P. (2020). *Artifical intelligence: A modern approach.* Upper Saddle River, NJ, US: Prentice Hall.

Sabin, S. (2019). *Voter support for facial recognition technology slips amid debates over its use.* Morning Consult. (Available at https://morningconsult.com/2019/06/19/voter-support-for-facial-recognition-technology-slips-amid-debates-over-its-use/ [accessed: 14 Jun. 2021])

Saha, D., Schumann, C., McElfresh, D. C., Dickerson, J. P., Mazurek, M. L., & Tschantz, M. C. (2020). Human comprehension of fairness in machine learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 152–152).

Salles, A., Evers, K., & Farisco, M. (2020). Anthropomorphism in AI. *AJOB neuroscience*, *11*(2), 88–95.

Shearer, E., & Grieco, E. (2019). *Americans are wary of the role social media sites play in delivering the news* (Tech. Rep.). Pew Research Center. Retrieved from `https://perma.cc/JLH2-C6H4`

Smith, A. (2016). *Public predictions for the future of workforce automation* (Tech. Rep.). Pew Research Center. Retrieved from `https://www.pewinternet.org/2016/03/10/public-predictions-for-the-future-of-workforce-automation/`

Smith, A. (2018). *Public attitudes toward technology companies* (Tech. Rep.). Pew Research Center. Retrieved from `https://perma.cc/96GP-7VFS`

Smith, A. (2019). *More than half of U.S. adults trust law enforcement to use facial recognition responsibly* (Tech. Rep.). Pew Research Center. Retrieved from `https://perma.cc/FUV7-5BDJ`

Smith, A., & Anderson, M. (2016). *Automation in everyday life* (Tech. Rep.). Pew Research Center. Retrieved from `https://perma.cc/WU6B-63PZ`

Spinks, C. N. (2019). Contemporary housing discrimination: Facebook, targeted advertising, and the fair housing act. *Houston Law Review*, *57*, 925.

Stark, B., Stegmann, D., Magin, M., & Jürgens, P. (n.d.). *Are algorithms a threat to democracy? the rise of intermediaries: A challenge for public discourse* (Tech. Rep.). AlgorithmWatch.

Swart, J. (2021). Experiencing algorithms: How young people understand, feel about, and engage with algorithmic news selection on social media. *Social Media+ Society*, *7*(2), 20563051211008828.

Thewissen, S., & Rueda, D. (2019). Automation and the welfare state: Technological change as a determinant of redistribution preferences. *Comparative Political Studies*, *52*(2), 171–208.

Thiebes, S., Lins, S., & Sunyaev, A. (2020). Trustworthy artificial intelligence. *Electronic Markets*, 1–18.

Vincent, J. (2021, jun). *Have autonomous robots started killing in war?* The Verge. (Available at: https://www.theverge.com/2021/6/3/22462840/killer-robot-autonomous-drone-attack-libya-un-report-context [accessed: 14 Jun. 2021])

Warren, A., & Hillas, A. (2020). Friend or frenemy? the role of trust in human-machine teaming and lethal autonomous weapons systems. *Small Wars & Insurgencies*, *31*(4), 822–850.

West, S. M., Whittaker, M., & Crawford, K. (2018). *Discriminating systems: Gender, race, and power in AI* (Tech. Rep.). AI New Institute. Retrieved from `https://ainowinstitute.org/discriminatingsystems.pdf`

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The role and limits of principles in AI ethics: towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 195–200).

Wu, N. (2019). *Misattributed blame? attitudes towards globalization in the age of automation.* (Working Paper)

Young, K. L., & Carpenter, C. (2018). Does science fiction affect political fact? yes and no: A survey experiment on "killer robots". *International Studies Quarterly*, *62*(3), 562–576.

Zhang, B. (2019). *No rage against the machines: Threat of automation does not change policy preferences.* (Working paper)

Zhang, B., Anderljung, M., Kahn, L., Dreksler, N., Horowitz, M. C., & Dafoe, A. (2021). Ethics and governance of artificial intelligence: Evidence from a survey of machine learning researchers. *Journal of Artifical Intelligence Research*, *71*, 591—666.

Zhang, B., & Dafoe, A. (2019). *Artificial intelligence: American attitudes and trends* (Tech. Rep.). Centre for the Governance of AI, University of Oxford. Retrieved from `http://dx.doi.org/10.2139/ssrn.3312874`

Zhang, B., & Dafoe, A. (2020). US public opinion on the governance of artificial intelligence. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (p. 187–193). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3375627.3375827` doi: 10.1145/3375627.3375827

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power.* New York: PublicAffairs.