Public Preferences for Governing AI Technology: Comparative Evidence

Sönke Ehret

The Version of Record of this manuscript has been published and is available in the Journal of European Public Policy 22 Aug 2022 DOI: 10.1080/13501763.2022.2094988

Abstract

Citizens' attitudes concerning aspects of AI such as transparency, privacy, and discrimination have received considerable attention. However, it is an open question to what extent economic consequences affect preferences for public policies governing AI. When does the public demand imposing restrictions on – even or prohibiting emerging AI technologies? Do average citizens' preferences depend causally on normative and economic concerns or only one of these causes? If both, how might economic risks and opportunities interact with assessments based on normative factors? And to what extent does the balance between the two kinds of concerns vary by context? I answer these questions using a comparative conjoint survey experiment conducted in Germany, the United Kingdom, India, Chile, and China. The data analysis suggests strong effects regarding AI systems' economic and normative attributes. Moreover, I find considerable cross-country variation in normative preferences regarding the prohibition of AI systems vis-a-vis economic concerns.

Keywords: AI Ethics, AI Governance, Automation Threat, Conjoint

Introduction

The idea to restrict harmful applications of Artificial Intelligence (AI)[1] has intuitive

[1] For a definition,I follow Nitzberg & Zysman in this special issue: "AI is *technology that uses advanced computation to perform at human cognitive capacity in some task area*" (Nitzberg & Zysman 2021, 4). This definition is closely related to Büthe (introduction to this special issue). Specifically, in this study I also refer to "AI systems", which "[…] carry out a wide variety of tasks with some degree of autonomy, i.e., without simultaneous, ongoing human intervention. Their capacity for learning allows AI systems to solve problems and thus support, emulate or even improve upon human decisionmaking […]" (Büthe, 2022, introduction to this special issue). AI systems thus possess the growing potential to perform both routine and non-routine, complex tasks (OECD, 2019, Ch.3).

appeal, culminating in the proposal to "blocklist" certain AI (European Commission, 2021). Experts are thus developing a regulatory agenda with far-reaching implications. However, little is known about the public's preferences, which will experience the consequences of AI systems. In particular, researchers have paid little attention to whether the public places equal emphasis on the economic and especially the labor market consequences as on the ethical implications of introducing AI.

Prior research has identified two ways the public forms preferences regarding AI systems. The first strand of research emphasizes AI systems' societal and ethical attributes. In this view, citizens are motivated by normative concerns for the governance of AI. For example, if AI systems harm society, discriminate against minorities, or breach common norms of transparency, placing restrictions on their deployment is strongly preferred by citizens (Cave & Dihal, 2019; Eurobarometer, 2017; Smith & Anderson, 2017). The second strand of literature emphasizes the various political consequences of labor-replacing technologies. This research has claimed that economic incentives strongly motivate public policy demands. It argues that automation-led unemployment leads to a backlash in the form of populism, less trust in government (Frey et al., 2018; Iversen & Soskice, 2019, Ch. 5; Kurer & Palier, 2020), and increasing the demand for redistribution (Thewissen & Rueda, 2019). Nevertheless, this literature has paid less attention to possibly distinctive assessments of technological change regarding AI.

The scope of political responses to AI remains an open question. When do citizens prefer regulatory protections? Many AI applications have tangible economic consequences and depend on normatively problematic features (Kellogg et al., 2020; OECD, 2019). However, no empirical study has identified whether preferences to prohibit AI systems depend causally on both or only one of these causes. Moreover,

doing so is methodologically challenging. Labor market and normative attributes may weigh heavily on citizen preferences, but whether one confounds the other is an open question. For instance, an AI system leading to unemployment may also be discriminatory, particularly for those losing their jobs.

I use conjoint experiments to answer these questions[2]. This survey experiment design allows me to test several causal hypotheses simultaneously (Bansak et al., 2018; Hainmueller & Hopkins, 2014). The analysis uses samples from Germany, the United Kingdom, Chile, China, and India. This case selection is based on a most different cases design logic, aiming to assess the impact of normative and economic attributes in various contexts.

Following this introduction, I first locate my contribution to current research on citizen preferences for the public management of AI. I proceed then to a discussion of my research approach and sample. A presentation of the results follows this discussion. I expand my analysis by providing predicted choice probabilities and conclude with a discussion on understanding citizen preferences in global AI governance.

Towards understanding citizen preferences on AI policy governance

A growing body of research is currently being conducted on the general public's sentiments on AI in the US and Europe. It has found a rising awareness of AI as an issue (Edelman, 2019; Fast & Horvitz, 2016; Zhang & Dafoe, 2019), with positive sentiments (Fast & Horvitz, 2016) but also pessimism (Cave et al., 2019; Kelley et al., 2021) about human rights (Eurobarometer, 2017), human control (Araujo et al., 2020;

[2] This type of experiment usually involves a choice between two alternatives with multiple, randomized characteristics.

Dietvorst et al., 2015; Johnson & Verdicchio, 2017) and unemployment (McClure, 2018; Smith & Anderson, 2017). Popular narratives about killer robots or bio-metric mass surveillance (Cave et al., 2019) underline the public's concerns about AI systems. However, current debates on the future shape of AI governance take place predominantly among experts. They can be characterized as an attempt to find general principles conforming to the law and societal values (European Commission, 2019; Krafft et al., 2020; Yeung, 2018).

Prior research has found that 82% of Americans (Zhang & Dafoe, 2019) and 88% of European respondents (Eurobarometer, 2017) prefer careful management of AI. However, the motivations behind such citizen preferences should be scrutinized further. It is a general question of whether policy issue attitudes are motivated by narrow economic interests, defined in terms of economic returns gained for oneself or a group (Scheve & Slaughter, 2001; Weeden & Kurzban, 2017), or whether normative, cultural, or similar societal concerns play a more important or rivalrous role than material motives (Fehr & Falk, 2002). Overall this research does not conclusively support narrow material interests as a significant driver of policy preferences (Lau & Heldman, 2009; Sears & Funk, 1990). Therefore, regarding AI systems, two strands of the literature suggest the possibility of either predominantly normative or economic motivations.

Some AI systems are met by strong normative expectations of the public (Araujo et al., 2020; Bigman & Gray, 2018; Shin, 2021). There is a general tendency of individuals to distrust algorithms (Burton et al., 2020; Dietvorst et al., 2015; Logg et al., 2019), conditional on the expectations placed on algorithms in their domains (Castelo et al., 2019; Logg et al., 2019; Prahl & Van Swol, 2017). Individuals evaluate the acceptability of AI systems in terms of moral costs and benefits (Kodapanakkal et al.,

2020). However, there is scarce causal evidence of how these trust-related algorithms' properties relate to preferences prohibiting specific AI systems. Except for Kodapanakkal et al. (2020), prior research has predominantly focused on a single category of application (Kelley et al., 2021).

My investigation confronts some of the predictions of this literature with an additional factor - the labor market costs and benefits of AI systems. There are multiple theoretical reasons why economic gains or losses would matter. There are direct incentives from income and employment (Becker, 1983; Meltzer & Richard, 1981) and indirect ones, most importantly, the loss of value of specific individual human capital (Frey & Osborne, 2017, 39; Gallego & Kurer, 2022; Iversen & Soskice, 2019, Ch. 6.2.1) due to AI. Research that individual economic displacement can be associated with political and populist backlash and analogous demands for public policy supports these insights (Frey et al., 2018; Im et al., 2020; Thewissen & Rueda, 2019).

Lastly, the present research is strongly focused on the United States and Europe. Among other parts of the world, such as India, or South-East Asia, evidence on public policy preferences relating to AI and AI systems is scant (Kelley et al., 2021). It is an open question how much Western Euro-centric concerns apply to other populations. My study speaks to this gap by incorporating and comparing samples from different regions.

Choosing between algorithms

I measure preferences for AI systems as preferences on both the outcomes and properties of applications of AI.[3] The experiment is framed in terms of "algorithms" as a

[3] Imagine a workplace AI application. Version 1 results in productivity gains, leading to more hiring; it also reports the worker's activity to third parties without consent. Version 2 reports the same information but with prior consent; it leads to similar productivity gains, but those

shorthand for all sorts of AI systems. The primary variable of interest is the decision as to which algorithm out of two should be prohibited in its deployment (asking the respondent "which of the [two] algorithms should be prohibited"). This setup mirrors the idea that several complex AI systems have multiple considerations that present themselves to individuals simultaneously (Shadbolt et al., 2016). Respondents thus place strict constraints on one algorithm by making a choice, implicitly permitting the other.

The choice between prohibiting algorithms is a simplification. Government regulation is often nuanced, yet the experiment's prohibition framing reflects current debates about which type of AI should be restricted (European Commission, 2021). The framing has further benefits. It provides an unambiguous interpretation of the decision problem across the examined countries and encourages respondents to pay attention to well-defined trade-offs. As a drawback, respondents cannot make more nuanced decisions with these binary choices. To measure their additional preferences, I ask respondents to rate seven-point Likert scale items on approving the two presented algorithms.

In the following section, specific economic and normative concerns are discussed for the choice between algorithms, which ultimately inform the design of this experiment.

Economic and Normative Concerns

Unemployment risks due to AI emerge in several domains beyond the automatization of routine tasks (Frey & Osborne, 2017, 3). Previous research finds weak evidence for the

gains lead the employer to replace the worker. While both scenarios concur with common findings in research on the future of work (MacCarthy, 2019; OECD, 2019, Ch. 3), the preference of average citizens is unknown.

effect of objective automation risks on redistribution and welfare policy demands (Gallego et al., 2022; Zhang, 2019) or the importance of AI-related technological unemployment in the opinions of survey respondents compared to the other consequences of AI (Zhang & Dafoe, 2019, p. 18). Moreover, the existence of a dominant effect of economic insecurity on backlash has been placed in doubt (Margalit, 2019). Therefore, the possibility of technological unemployment risks with four different degrees of severity is incorporated. My design makes the unemployment risk blunt for the respondent, stating that, in the extreme case, the respondent and others like the respondent would face a significant risk of unemployment due to automation, indicating the labor-replacing effect of the AI system in question. This prompt removes the cognitively demanding uncertainty, often inherent in predictions linked to individual circumstances, e.g., skill profiles.

One obvious countervailing argument to the technological unemployment risk is the potential economic benefits. The net macroeconomic effect of AI is ambiguous since new occupations and jobs are created, and productivity is enhanced (Autor, 2015). Given these benefits, citizens face a trade-off. One option is trying to curtail algorithms that displace jobs generally, but, at the same time, one can accept job displacement for the promise of new economic opportunity. The design incorporates this dimension by including the possibility of positive economic consequences of AI and specifying which occupational group (labor intensive, capital intensive, or both) benefits. While there is a strong focus on technological unemployment, AI's economic benefits have not been reflected much in prior empirical studies on public policy preferences. Still, they are important, given the role material incentives can play in forming public policy preferences.

These countervailing possible economic consequences lead to the first two

hypotheses:

H_{1A}. If an algorithm increases the risk of unemployment, respondents will be more likely to prohibit it.

H_{1B}. If an algorithm increases employment gains, respondents will be less likely to prohibit it.

Beyond their economic consequences, AI systems also might emulate human choices and judge normatively complex situations (Garfinkel et al., 2017; Veale et al., 2018). There is currently scant causal evidence on the effects of these normative AI features on preferences prohibiting specific AI systems. Therefore, two important categories of decision making, discrimination and making distributive choices, are considered. Discrimination is an important issue (Caliskan et al., 2017; Veale et al., 2018), and prior surveys point to citizens disliking unfair algorithms (Binns et al., 2018). An item is included in this study that presents various degrees of discrimination.

Since AI systems make choices, these systems' trustworthiness and morality are often explicitly evaluated. Prior research has identified empirical support for the thesis that more explainable algorithms are trusted more (Burton et al., 2020; Shadbolt et al., 2016; Shin, 2021). Therefore, simple descriptive attributes are included to test explainability in three levels, ranging from easy to explain to impossible to understand.[4] Furthermore, individuals commonly require moral competency for critical distributive decisions and morally judge AI systems making such decisions (Binns et al., 2018), often with skepticism against AI (Bigman & Gray, 2018). I expect that an AI system will be trusted more and requested to be prohibited less if critical choices are made between things rather than between humans and things. Three dimensions of choice-

[4] Going forward explainability will be interpreted through its result, transparency.

making between humans and things are incorporated to test this proposition.

The degree of user privacy is also incorporated. There is preliminary evidence linking privacy violations to demands for tighter regulations on all sorts of computerized applications (Kodapanakkal et al., 2020; Nissenbaum, 2018). An analogous preference of individuals is expected, i.e., to prohibit systems that collect user data without consent, rewarding data collection with consent, or the ability to customize which data is collected. Furthermore, the algorithm's deployment domain is included in the design as an experimental control. An algorithm that makes unfair decisions in the justice and law field is different from systems deployed in factories[5] (Büthe, 2022; Nitzberg & Zysman, 2021).

H_2. Faced with normatively objectionable AI features, respondents will be more likely to prohibit an algorithm.

The conjoint design allows me to combine the abovementioned normative and economic concerns into a testable framework. Seven randomized attributes are incorporated: i) technological unemployment risk, ii) job creation, iii) privacy concerns, iv) transparency, v) discrimination, vi) domain of application, and vii) distributive choices. Table 1 in the Online Appendix provides further details on the design.

Exploring Differences across Countries

Germany, the United Kingdom, Chile, and China differ on the level of economic development, the provision of public goods, and the welfare state, the policies that insulate an individual from economic change. Furthermore, there are cultural differences, different experiences with algorithms in daily life, prior abuses of

[5] I am agnostic about the precise consequences of the application domain for prohibition preferences.

individual rights, and political representation.

Given the differences in the sample, it would be plausible to expect variation. On the other hand, there are limits to making generalizable statements with small samples (Henrich et al., 2010), and accounts of the degree of similarities of normative behavior across countries and cultures are ambiguous (Henrich et al., 2010; Oosterbeek et al., 2004). Furthermore, a host of evidence points to universal material and non-material human motivations (Fehr & Falk, 2002). Given the precise experimental framing for the effect of a single experimental attribute across countries, I expect to find similar individual reactions.

A second concern pertains to a substantive interpretation of the choices made. Expectations are drawn from common theories about cross-country differences in the balance between material and non-material values (Inglehart, 2008). They predict that individuals in post-industrial societies place greater weight on non-material than material values. Accordingly, due to the level of economic development, individuals should pay more attention to normatively desirable AI systems in countries such as the United Kingdom and Germany.[6] This explanation is used as a plausible benchmark to acknowledge multiple potential other explanations for cross-national variance.[7] To describe this variation, selected combinations of AI attributes are considered and computed across countries, and probabilities for the choice are predicted in favor of the "prohibition" of specific cases of AI systems.

H3. Respondents in more developed countries will place less weight on prohibiting AI systems with high unemployment risk relative to the normative

[6] A strict ordering using the post-materialism index by Inglehardt, based on data from 1970 to 2006 would yield the following order ranging from higher to lower values on the index: UK, Germany, Chile, India, China.

[7] For example, variation in socially consequential algorithms, such as China's social credit system, or in general attitudes regarding government regulation.

requirements of the AI system than respondents in less developed countries.

Sample

The experiment is part of a multi-investigator time-sharing experiments project conducted in 2019 at a large UK-based public research university. Data was collected in 2019 for the United Kingdom, Chile, and India. This data has been obtained from the UK university's online subject opt-in panels (in 2019, 8,000 subjects in the UK, 10,000 in Chile, and 20,000 in India). In mid-2019, additional data was collected in China and Germany. Samples have been obtained via large online sample platforms Clickworkers (Germany) and Microworkers (China), which are also online opt-in panel providers. Participants have been remunerated with a flat fee commensurate with the country and survey time. The Online Appendix, Samples section provides more details on the number of observations, the sample source, and data collection.

In total, 932 respondents were recruited in Germany, the UK, India, Chile, and China. Respondents are, on average middle-aged, as in Germany (M: 39.4), the UK (M: 42.9), Chile (M: 47.0), and India (M: 27.9). Except for India, the samples are balanced in gender: 48% female in Germany, 51% female in the UK, 52% female in Chile, and 30% female in India. The sample in India is biased towards urban, educated, primarily male subjects. More details on the demographic background can be found in table 1.

Table 1: Key demographics of the German, UK, Chile, and Indian samples. Data for China is unavailable.

Country	Age (years)	Gender (0:M, 1:F)	Highest degree obtained	Household income
	mean, sd	mean, sd	mode	mode
UK	42.9, 14.3	0.51, 0.50	Bachelor's degree	20,000 - 39,999 GBP
Germany	39.4, 12.4	0.48, 0.50	High school or equivalent	20.000 - 39.999 Euro
Chile	47.0, 13.9	0.52, 0.50	Bachelor's degree	448,001 to 1,000,000 CLP
India	27.5, 8.9	0.30, 0.46	Bachelor's degree	600,000 Rs and over

There are similar levels of employment among those who answered the survey question on employment: 59% in Germany, 58% in the UK, 56% in China, 55% in Chile, and 43% in India. Figure 4 in the Online Appendix shows the distributions of self-reported occupations in the samples. For all countries, there is an intense concentration of professions requiring training or education (professional, managerial, administrative, and technical), mainly in non-manual occupations. Overall, there is a considerable similarity between the samples regarding basic occupational background. There are also comparable distributions regarding self-reported AI knowledge; see Online Appendix, figures 7 and 8.

Measurements and Results

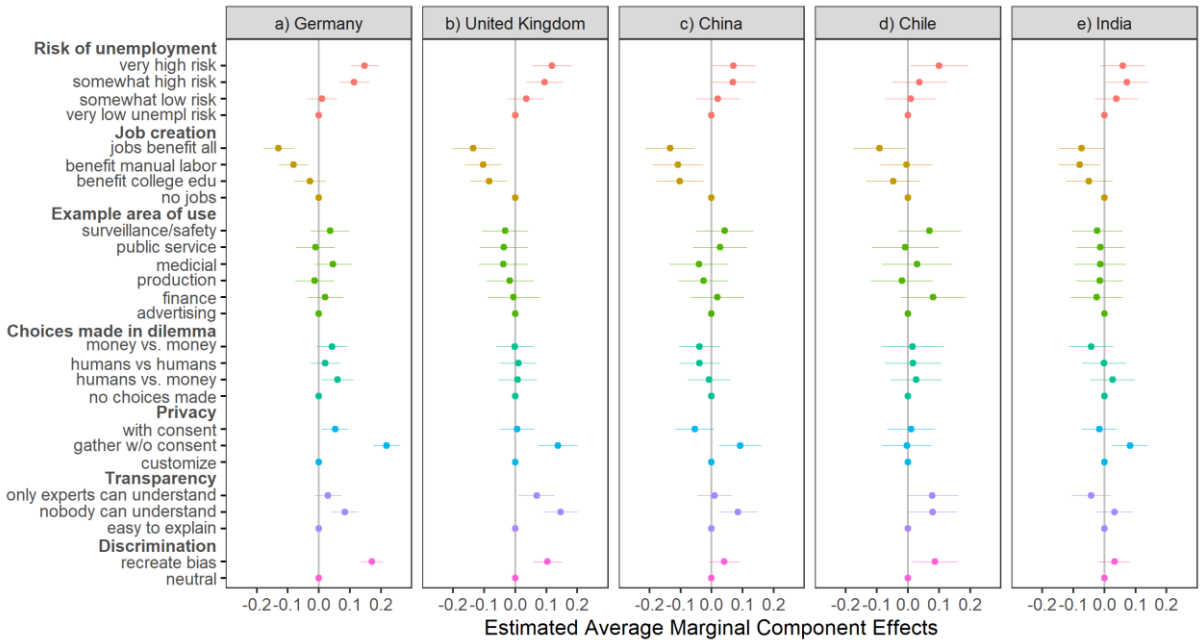
Each respondent faced five algorithm choice trials, during which two randomly chosen algorithm profiles were presented side-by-side in tabular format, resulting in 9320 subject-trial-profile observations. Across respondents, the attributes of the algorithm

profiles were presented in randomized order. However, the order was kept fixed for a given respondent to ensure that the choice sets for a given respondent differed only concerning the substantive attributes.

For each pair, the respondent had to choose which algorithm to prohibit. In addition to this binary measure, a seven-point Likert scale response was collected on approving rather than restricting each of the two algorithm profiles. The two measures are negatively correlated. We present results based on the Likert scale measure in the Online Appendix, figure 2.

The analyses reveal several effects that are common across all countries. However, differences across countries are also found.

Figure 1: Estimated Average Marginal Component Effects



The average marginal component effect is estimated, estimating the average change of the probability that the respondent prefers a particular algorithm to be prohibited, comparing the values of a given attribute versus a baseline value (Bansak et al., 2018; Hainmueller & Hopkins, 2014), for each country in the sample. The estimated effect can be interpreted as the treatment effect averaged across the distribution of all possible combinations of the attributes.[8] The estimated model also includes non-experimental control covariates on individual expectations of living standards, dependency on unemployment benefits, concerns about privacy intrusion, and knowledge items on digital literacy. The question wordings can be found in the Online Appendix.

Figure 1 shows the effect plots for an AMCE model estimated via linear regression. Solid dots represent point estimates; the lines indicate 95% confidence intervals. The omitted reference value is shown as a dot at the bottom of each attribute coefficient list. Standard errors are clustered on the respondent level since each respondent makes correlated choices within the five trials.[9]

The estimate indicates an increase (positive values) or a decrease (negative values) in the probability that a respondent chooses to prohibit an algorithm with a given attribute compared to an algorithm with the omitted baseline attribute. Persistent regularities are found across all countries. Regarding the labor market, respondents are

[8] All attributes were uniformly distributed in this study, since the attribute value combinations are not mutually exclusive.

[9] The model is estimated for the fully pooled sample and assessed whether there are profile order effects. (Figure 6 in the Online Appendix).

prompted with a variable risk of losing their job due to introducing a particular AI application. The baseline (omitted category) is facing no such unemployment risk. Compared to facing no such risk, facing high risks of unemployment led respondents to prefer prohibiting these particular algorithms increasingly. In Germany this increased prohibition probability is 15% ($p < 0.01$), in the UK it is 11% ($p < 0.01$), in China 7 % ($p < 0.05$), Chile 11% ($p < 0.05$), and in India (for somewhat high risks of unemployment) 7% ($p < 0.05$), compared to a situation where no such job loss looms. The employment effect matters for the propensity to prohibit a specific algorithm. For all countries, it is found that high unemployment risk has a positive, statistically significant effect on prohibition preferences (H1A).

Economics also matters for the positive consequences of AI. If there is job creation, respondents are less likely to request the prohibition of an AI system, holding other factors constant. Again, this pattern holds across all surveyed countries. Looking at an algorithm that creates new jobs, in Germany, the probability of prohibiting such an algorithm decreases by 13% ($p < 0.01$) and in the UK by 14% ($p < 0.01$). It decreases the probability of prohibition in China by 13% ($p < 0.01$), in India 7% ($p < 0.1$), and in Chile by 9% ($p < 0.05$), compared to a case where no new jobs are created.

These labor market considerations matter even when considering the non-economic aspects of transparency, distributive choices, application domain, discrimination, and privacy. Despite the presence of normative considerations on the regulation of AI, respondents show concerns about its negative employment impact. However, such concern is only one component of a more balanced picture since countervailing economic benefits outweigh the push to prohibit an AI system. This finding is interesting since the jobs created in the conjoint scenario accrue to society as a whole, while the unemployment would hit people like the respondent directly. In

sum, hypotheses 1A and 1B are supported by evidence. I reject the conjecture that non-economic regulatory concerns, as listed above, have the power to confound economic concerns. It is not normative concerns alone that drive preferences to restrict new AI technology.

Regarding the normative attributes, significant effects are found, too, implying that even when controlling for pertinent employment effects, normative considerations remain important (H2), but to varying degree depending on the country and issue. Markedly, these attributes exhibit greater cross-country variation than the labor market implications. Important topics such as privacy violations find higher a propensity to prohibit in Germany (+22 %, $p < 0.01$) and the UK (+14 %, $p < 0.01$), India (+8 %, $p < 0.01$), China (+9 %, $p < 0.1$) but not in Chile (-3 %, $p = 0.95$). Conversely, discriminatory bias substantially impacts prohibition preferences in Germany (17 %, $p < 0.01$), the UK (+10 %, $p < 0.01$) and Chile (+9 %, $p < 0.05$), but less so in China (+4 %, $p = 0.11$) or India (+2 %, $p = 0.2$). A similar pattern can be seen regarding the transparency of algorithms: The German, UK and Chilean respondents show higher aversion to difficult-to-understand algorithms than respondents in China or India.

Intuitively, one would expect, for example, that a system deployed for surveillance that makes choices between humans and things would be viewed critically in the eyes of an average individual. However, these effects appear to be relatively weak. No significant evidence on the remaining decision-making and domain choices is found.

Figure 1 in the Online Appendix shows the differences in the estimated effects across countries, taking Germany as a reference. Interestingly, no strong evidence exists for cross-country variation for job creation effects regarding the average response to labor market factors. Furthermore, comparing Germany to India, Chile, and China

(or the UK), there are only limited statistically significant differences in prohibition preferences in the face of unemployment risk.[10] German respondents appear to demand slightly more unemployment risk-related prohibition than respondents in India ($p < 0.05$) or China ($p < 0.1$).

For three normative issue variables – privacy, discrimination, and transparency – cross-country differences were found at varying degrees of magnitude across countries. Gathering data without consent is more likely to lead to prohibition demands in Germany than in Chile (22 percentage point difference, $p < 0.01$), India (14 percentage point difference, $p < 0.01$), or China (13 percentage point difference, $p < 0.01$). The UK has a comparable lead over Chile, India, and China. A similar pattern emerges for discrimination and, to some extent, transparency. This finding somewhat supports the notion that in wealthier countries, the average respondent pays greater attention to specific normative features of AI systems (H3). When considering the labor market dimension, there is similarity and variation between countries, depending on the nature of the attributes.

The absolute and difference AMCE estimations are conducted as a final robustness check with the alternative ordinal (Likert scale-based) dependent variable. This variable measures the approval of an algorithm without forcing a choice. Results can be found in figure 2 (Online Appendix). The findings of this analysis closely mirror the results from above. Across countries, respondents approve of algorithms that create jobs and disapprove of algorithms associated with high unemployment risk. In addition, there is significant variation in approval depending on the ethical/social attribute across countries.

[10] The finding does not rule out the possibility of differences of smaller magnitude.

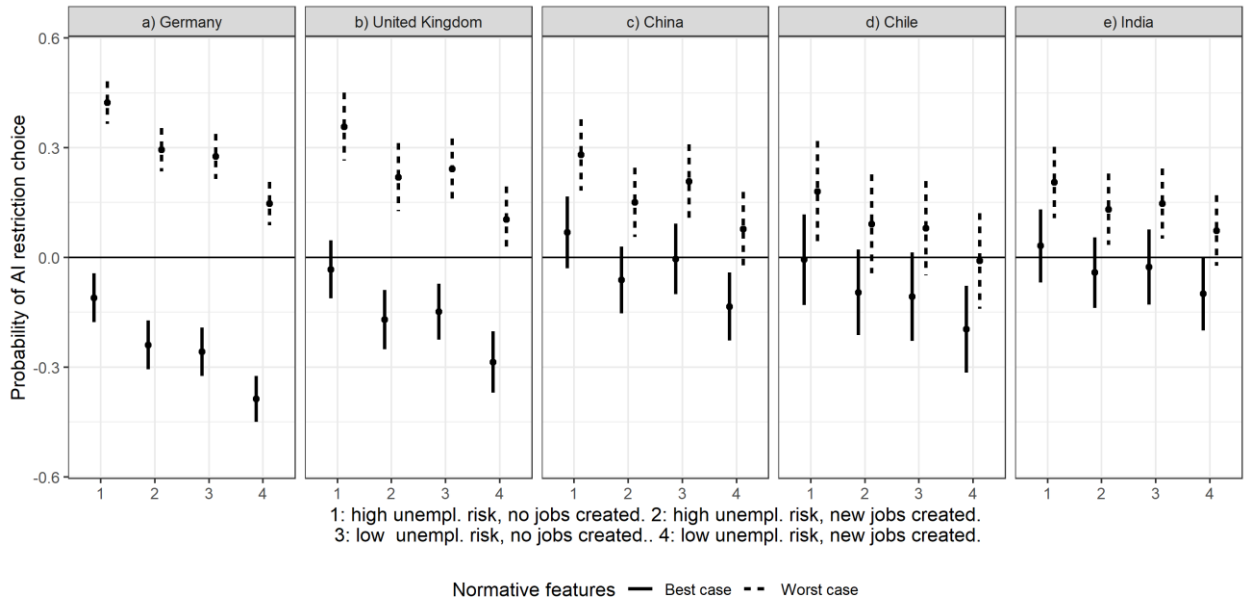
Predicted Prohibition Probabilities

Predicted prohibition propensities are calculated based on select attribute combinations, which vary across two dimensions. For the first dimension, there is a normatively optimal scenario ("best-case") and a scenario where harmful AI systems take place ("worst-case"). In the "worst-case" scenario, an algorithm submits private user data without consent. It is discriminatory, makes judgment choices between humans and money, and cannot be understood by anybody. This worst-case is contrasted with an algorithm where rather desirable outcomes take place. Here, the user can control the data submitted by the algorithm. It is socially neutral, does not make any judgments between humans and money, and can be easily understood.

For the second dimension, the economic consequences are varied. Four scenarios are considered. i) No new jobs are created, but a high unemployment risk exists. This is the worst-case scenario (Frey & Osborne, 2017, 39-42). ii) New jobs are created, but unemployment risks are high (Summers, 2013). iii) Low unemployment risk, but no new jobs created (Gordon, 2018). iv) Low unemployment risk and new jobs (Autor, 2015).

Those are eight profiles for which the predicted prohibition probabilities are computed for each of the five countries using a linear model. Figure 2 shows the plotted coefficients and 95% confidence intervals for the demeaned predicted choices, clustering on the respondent level. A coefficient of zero in the following analysis reflects the average choice of prohibition. For absolute values, see figure 3, Online Appendix.

Figure 2: Predicted choice probabilities for selected profiles are based on an OLS estimation regressing the conjoint choice on the conjoint attributes (respondent-level clustered standard errors). Profiles relate to a combination of the extreme outcomes for unemployment and the normative attributes explained in the text (domain of application was not specified). Confidence intervals were calculated using the delta method described in Xu and Long (2005) and implemented in the *prediction* package in R.



The plot in figure 2 shows remarkable patterns. Normatively desirable AI systems decrease the overall willingness to demand the prohibition of the introduction of AI. This pattern can be seen in the case of Germany and the UK, and to some extent, in China. The effects of normatively desirable AI systems are strong for Germany and the UK. Even when net unemployment emerges (Fig.2, scenario 1), there is no strong tendency to demand restrictions above the average. In this case, a normatively desirable AI system is associated with prohibition preference below the average (Germany) or preferences indistinguishable from the average (the UK, Chile, India, and China).

Irrespective of the economic consequences, an algorithm will be more likely to be subject to prohibition if it has normatively undesirable consequences. Even if no unemployment risk exists and new jobs are generated for capital and labor-intensive occupations (Fig.2, scenario 4), there is no evidence suggesting that less than average restrictions would be imposed. Despite the net gain implied, there is sometimes an above-average increase in predicted restriction decisions (Germany: $p < 0.05$, UK: $p < 0.1$). There is a weaker pattern for questionable AI systems in China, India, and Chile. The negative impact of ethically and socially uncertain AI systems vastly outweighs economic net benefits.

Finally, countries with at least some but not full economic benefits conditional on AI are considered in the best-case scenario. Here the greatest variation across countries is found. Germany and the UK are analyzed in the beginning. Here, high unemployment risk coupled with job creation, or low unemployment risk without new jobs, is associated with a decrease in prohibition demands below average ($p < 0.01$). Only the case of full economic benefits in India and China appears to move respondents into a decrease in prohibition demands below average ($p < 0.05$). In Chile, however, intermediate cases have an effect, yet to a less reliable extent ($p < 0.1$).

This variation underlines the exploratory conjecture (H_3). In more developed countries, respondents put greater weight on normative aspects of AI systems, in contrast to the more uniformly significant labor market consequences of these systems.

Conclusion

Discussions over governing AI often have a predominantly normative focus. Evidence about the extent to which those concerns reflect the preferences of individual citizens is still quite limited. The experimental research presented in this paper provides such

evidence across five countries from different parts of the world. It investigates an under-researched question: Are citizens' preferences only shaped by the normative aspects of AI systems, or do material consequences also play an essential role?

The results point to both material and normative concerns. On average, individuals would opt for banning systems that cause unemployment once normative considerations, application domains, and additional job creation benefits are considered. However, I also find that normatively robust AI applications have a very significant impact on individual preferences.

For well-known domains such as immigration, trade, and environmental protection, similar "mixed" motives are important (Bechtel et al., 2019). Compared to these prominent political issues, my findings suggest that material and normative concerns already resonate in respondents' minds for a novel policy domain such as AI. More research is needed, however, to parse out further why this is the case.

One reason for my findings is a more general preference to seek protection from technological unemployment (Desmet & Parente, 2014; Frey et al., 2018; Gallego et al., 2022). Evidence for direct policy preferences for protection from technological unemployment is scarce, partly due to confounding with other preferences citizens have. The design of my study controls these issues experimentally, suggesting that citizens might prefer to use regulatory power to prohibit certain types of AI. Thus, my study finds that there are direct consequences of technological unemployment risk for policy preferences, in addition to previously identified indirect political effects of such risk (Im et al., 2020; König & Wenzelburger, 2019).

Moreover, I find that normatively well-regarded AI systems can receive a significant degree of support, more so in the European than non-European context, despite the risks they entail for employment. For example, general technology

restriction preferences appear to be predominant even in comparison to a preference for government redistribution (Gallego et al., 2022). My findings thus underline the need to examine further the role of social norms in the public acceptability of AI-related economic change and, ultimately, the legitimacy of specific approaches to AI governance.

More research is needed to understand the cross-country variation of normative preferences as they relate to the role of AI. Beyond post-materialism, there are broader, conflicting frameworks available. While extant research highlights the importance of cultural differences for value-based preferences (Hofstede, 2001), there is also important work on institutions and the cultural consequences of economic development (Bowles, 1998). Future work could further address these nuances. Hofstede emphasizes the cultural dimension of uncertainty avoidance, where Germany ranks high. Other countries, such as China or India, may be culturally associated with more technological risk-taking. My study gives the first empirical insight into the nature of such cross-country variation regarding AI.

I furthermore reanalyzed the data and re-coded professional backgrounds according to the skills required by the respondents' profession (see: Gallego & Kurer, 2022). For the results, see figure 5 in the Online Appendix. Skills do not modify the treatment effects for the economic attributes. A further investigation is needed relating to the degree to which individuals perceive themselves at risk from AI systems relating to their skills, education, and normative beliefs.

My study contributes to the debate on the feasibility of global AI governance (Cath et al., 2017; Nitzberg & Zysman, 2021; Schiff et al., 2020) based on a novel experiment on public policy preferences. The results of my study suggest that there is common ground regarding the economic consequences of AI governance. While the

predictability of the economic consequences of AI systems is an open question (Frank et al., 2019), by and large, individuals from very different countries recognize both the economic challenges and opportunities of AI. However, my findings suggest that European values regarding the restriction of certain AI systems may find less acceptance in other global publics. Whether global AI governance is shaped by its economic consequences or normative "logic of appropriateness" (March & Olsen, 1998, 951) should not be seen as a dichotomy. Integrating cultural and economic development options into global AI governance frameworks (Jobin et al., 2019) could create maximum public acceptance and regulatory flexibility.

Sönke Ehret is a PostDoc at the University of Lausanne.

Contact information:

of Lausanne, Faculty of Business and Economics, Quartier de Chamberonne, Internef, 1015 Lausanne. Email: sonkeklaus.ehret@unil.ch

Word count: 7936 words

Replication materials

Supporting data and materials for this article can be accessed at

https://github.com/soeehret/JEPP_replication_materials

Acknowledgements

I thank the special issue editor and the reviewers for their insightful, constructive comments and criticism. I also thank the organizers and participants of the

TUM Workshop on Governance of Artificial Intelligence for their valuable feedback. This research was fully or partly conducted through the Nuffield College Centre for Experimental Social Sciences (CESS). The authors gratefully acknowledge the support of CESS in conducting their research.

References

Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society*, *35*(3), 611–623. https://doi.org/10.1007/s00146-019-00931-w

Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives*, *29*(3), 3–30. https://doi.org/10.1257/jep.29.3.3

Bansak, K., Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2018). The number of choice tasks and survey satisficing in conjoint experiments. *Political Analysis*, *26*(1), 112–119. https://doi.org/10.1017/pan.2017.40

Bechtel, M. M., Genovese, F., & Scheve, K. F. (2019). Interests, norms and support for the provision of global public goods: The case of climate co-operation. *British Journal of Political Science*, *49*(4), 1333–1355. https://doi.org/10.1017/S0007123417000205

Becker, G. S. (1983). *Human capital: A theoretical and empirical analysis, with special reference to education* (2nd ed.). The University of Chicago Press.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). "It's reducing a human being to a percentage": Perceptions of justice in algorithmic decisions. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3173951

Bowles, S. (1998). Endogenous preferences: The cultural consequences of markets and other economic institutions. *Journal of Economic Literature*, *36*(1), 75–111. JSTOR.

Burton, J. W., Stein, M., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239. https://doi.org/10.1002/bdm.2155

Büthe, T. (2022). *Introduction to the Special Issue on AI Governance*.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2017). Artificial Intelligence and the 'Good Society': The US, EU, and UK approach. *Science and Engineering Ethics*, *24*, 505–528. https://doi.org/10.1007/s11948-017-9901-7

Cave, S., Coughlan, K., & Dihal, K. (2019). "Scary robots": Examining public responses to AI. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 331–337. https://doi.org/10.1145/3306618.3314232

Cave, S., & Dihal, K. (2019). Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence*, *1*(2), 74–78. https://doi.org/10.1038/s42256-019-0020-9

Desmet, K., & Parente, S. L. (2014). Resistance to technology adoption: The rise and decline of guilds. *Review of Economic Dynamics*, *17*(3), 437–458. https://doi.org/10.1016/j.red.2013.09.005

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Edelman. (2019). *2019 Edelman AI survey results report*. Edelman Research. https://www.edelman.com/sites/g/files/aatuss191/files/2019-03/2019_Edelman_AI_Survey_Whitepaper.pdf

Eurobarometer. (2017). *Attitudes towards the impact of digitisation and automation on daily life*. European Commission. https://europa.eu/eurobarometer/surveys/detail/2160

European Commission. (2019). *A definition of AI: Main capabilities and scientific disciplines*. https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf

European Commission. (2021). *Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts*. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN

Fast, E., & Horvitz, E. (2016). Long-term trends in the public perception of artificial intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*. https://doi.org/10.48550/arXiv.1609.04904

Fehr, E., & Falk, A. (2002). Psychological foundations of incentives. *European Economic Review*, *46*(4–5), 687–724. https://doi.org/10.1016/S0014-2921(01)00208-2

Frank, M. R., Autor, D., Bessen, J. E., Brynjolfsson, E., Cebrian, M., Deming, D. J., Feldman, M., Groh, M., Lobo, J., Moro, E., Wang, D., Youn, H., & Rahwan, I. (2019). Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences*, *116*(14), 6531–6539. https://doi.org/10.1073/pnas.1900949116

Frey, C. B., Berger, T., & Chen, C. (2018). Political machinery: Automation anxiety and the 2016 U.S. presidential election. *Oxford Review of Economic Policy*, *34*(3), 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. https://doi.org/10.1016/j.techfore.2016.08.019

Gallego, A., Kuo, A., Manzano, D., & Fernández-Albertos, J. (2022). Technological risk and policy preferences. *Comparative Political Studies*, *55*(1), 60–92. https://doi.org/10.1177/00104140211024290

Gallego, A., & Kurer, T. (2022). Automation, digitalization, and artificial intelligence in the workplace: Implications for political behavior. *Annual Review of Political Science*, *25*(1), 463–484. https://doi.org/10.1146/annurev-polisci-051120-104535

Garfinkel, S., Matthews, J., Shapiro, S. S., & Smith, J. M. (2017). Toward algorithmic transparency and accountability. *Communications of the ACM*, *60*(9), 5. https://doi.org/10.1145/3125780

Gordon, R. J. (2018). *Why has economic growth slowed when innovation appears to be accelerating?* (CEPR Discussion Papers No. 13039). National Bureau of

Economic Research. https://www.nber.org/system/files/working_papers/w24554/w24554.pdf

Hainmueller, J., & Hopkins, D. J. (2014). Public attitudes toward immigration. *Annual Review of Political Science*, *17*(1), 225–249. https://doi.org/10.1146/annurev-polisci-102512-194818

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nation* (2nd ed.). Sage.

Im, Z. J., Mayer, N., Palier, B., & Rovny, J. (2020). The "losers of automation": A reservoir of votes for the radical right? *Research & Politics*, *6*(1), 1–7. https://doi.org/10.1177/2053168018822395

Inglehart, R. F. (2008). Changing values among western publics from 1970 to 2006. *West European Politics*, *31*(1–2), 130–146. https://doi.org/10.1080/01402380701834747

Iversen, T., & Soskice, D. W. (2019). *Democracy and prosperity: Reinventing capitalism through a turbulent century*. Princeton University Press.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Johnson, D. G., & Verdicchio, M. (2017). Reframing AI discourse. *Minds and Machines*, *27*(4), 575–590. https://doi.org/10.1007/s11023-017-9417-6

Kelley, P. G., Yang, Y., Heldreth, C., Moessner, C., Sedley, A., Kramm, A., Newman, D. T., & Woodruff, A. (2021). Exciting, useful, worrying, futuristic: Public perception of artificial intelligence in 8 countries. *Proceedings of the 2021 AAAI/ACM*

Conference on AI, Ethics, and Society, 627–637. https://doi.org/10.1145/3461702.3462605

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals*, *14*(1), 366–410. https://doi.org/10.5465/annals.2018.0174

Kodapanakkal, R. I., Brandt, M. J., Kogler, C., & van Beest, I. (2020). Self-interest and data protection drive the adoption and moral acceptability of big data technologies: A conjoint analysis approach. *Computers in Human Behavior*, *108*, 106303. https://doi.org/10.1016/j.chb.2020.106303

König, P. D., & Wenzelburger, G. (2019). Why parties take up digitization in their manifestos: An empirical analysis of eight Western European economies. *Journal of European Public Policy*, *26*(11), 1678–1695. https://doi.org/10.1080/13501763.2018.1544268

Krafft, T. D., Zweig, K. A., & König, P. D. (2020). How to regulate algorithmic decision-making: A framework of regulatory requirements for different applications. *Regulation & Governance*, *16*, 119–136. https://doi.org/10.1111/rego.12369

Kurer, T., & Palier, B. (2020). Shrinking and shouting: The political revolt of the declining middle in times of employment polarization. *Research & Politics*, *6*(1), 2053168019831164. https://doi.org/10.1177/2053168019831164

Lau, R. R., & Heldman, C. (2009). Self-interest, symbolic attitudes, and support for public policy: A multilevel analysis. *Political Psychology*, *30*(4), 513–537. https://doi.org/10.1111/j.1467-9221.2009.00713.x

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

MacCarthy, M. (2019). How to address new privacy issues raised by artificial intelligence and machine learning. *Brookings Blog*.

https://www.brookings.edu/blog/techtank/2019/04/01/how-to-address-new-privacy-issues-raised-by-artificial-intelligence-and-machine-learning/

March, J. G., & Olsen, J. P. (1998). The institutional dynamics of international political orders. *International Organization*, *52*(4), 943–969. https://doi.org/10.1162/002081898550699

Margalit, Y. (2019). Economic insecurity and the causes of populism, reconsidered. *Journal of Economic Perspectives*, *33*(4), 152–170. https://doi.org/10.1257/jep.33.4.152

McClure, P. K. (2018). "You're fired," says the robot: The rise of automation in the workplace, technophobes, and fears of unemployment. *Social Science Computer Review*, *36*(2), 139–156. https://doi.org/10.1177/0894439317698637

Meltzer, A. H., & Richard, S. F. (1981). A rational theory of the size of government. *Journal of Political Economy*, *89*(5), 914–927.

Nissenbaum, H. (2018). Respecting context to protect privacy: Why meaning matters. *Science and Engineering Ethics*, *24*(3), 831–852. https://doi.org/10.1007/s11948-015-9674-9

Nitzberg, M., & Zysman, J. (2021). *Algorithms, data, and platforms: The diverse challenges of governing AI*. BRIE Working Paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3802088

OECD. (2019). *Artificial Intelligence in Society*. https://www.oecd-ilibrary.org/content/publication/eedfee77-en

Oosterbeek, H., Sloof, R., & van de Kuilen, G. (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics*, *7*(2), 171–188. https://doi.org/10.1023/B:EXEC.0000026978.14316.74

Prahl, A., & Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, *36*(6), 691–702. https://doi.org/10.1002/for.2464

Scheve, K. F., & Slaughter, M. J. (2001). What determines individual trade-policy preferences? *Journal of International Economics*, *54*(2), 267–292. https://doi.org/10.1016/S0022-1996(00)00094-5

Schiff, D., Biddle, J., Borenstein, J., & Laas, K. (2020). What's Next for AI Ethics, Policy, and Governance? A Global Overview. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 153–158. https://doi.org/10.1145/3375627.3375804

Sears, D. O., & Funk, C. L. (1990). The limited effect of economic self-interest on the political attitudes of the mass public. *Journal of Behavioral Economics*, *19*(3), 247–271. https://doi.org/10.1016/0090-5720(90)90030-B

Shadbolt, N., Van Kleek, M., & Binns, R. (2016). The rise of social machines: The development of a human/digital ecosystem. *IEEE Consumer Electronics Magazine*, *5*(2), 106–111. https://doi.org/10.1109/MCE.2016.2516179

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Smith, A., & Anderson, M. (2017). *Automation in everyday life*. Pew Research Center. http://www.pewinternet.org/2017/10/04/automation-in-everyday-life

Summers, L. H. (2013). *Economic possibilities for our children* (No. 4; The Reporter). National Bureau of Economic Research. https://www.nber.org/reporter/2013number4/economic-possibilities-our-children

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. https://doi.org/10.1177/0010414017740600

Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. *Proceedings of the 2018 Chi Conference on Human Factors in Computing Systems*, 1–14.

Weeden, J., & Kurzban, R. (2017). Self-interest is often a major determinant of issue attitudes: Self-interest and issue attitudes. *Political Psychology*, *38*, 67–90. https://doi.org/10.1111/pops.12392

Xu, J., & Long, J. S. (2005). Confidence intervals for predicted outcomes in regression models for categorical outcomes. *The Stata Journal*, *5*(4), 537–559. https://doi.org/10.1177/1536867X0500500405

Yeung, K. (2018). Algorithmic regulation: A critical interrogation: Algorithmic regulation. *Regulation & Governance*, *12*(4), 505–523. https://doi.org/10.1111/rego.12158

Zhang, B. (2019). *No rage against the machines: Threat of automation does not change policy preferences* (No. 3455501). SSRN Working Paper. https://doi.org/10.2139/ssrn.3455501

Zhang, B., & Dafoe, A. (2019). *Artificial intelligence: American attitudes and trends* (No. 3312874). SSRN Working Paper. https://doi.org/10.2139/ssrn.3312874