JOHN P. NELSON

What Do We Want From AI?

Public values must shape governance and implementation of artificial intelligence.

Should artificial intelligence systems diagnose diseases and prescribe treatments without human oversight? Should writers and artists have the option to bar their works from being used to train generative AI? Should firms be able to use algorithmic management systems to subconsciously "nudge" workers' behavior or to record and extract their knowledge? Should military drones be allowed to make the decision to fire independently?

The advance of AI spawns new political and ethical questions like these almost daily. These questions are debated by pundits, policy wonks, researchers, and business leaders interested in guiding AI development and implementation in responsible, ethical, and beneficial directions. Several scholars have argued that a consensus has emerged around a set of principles to guide AI ethics. These principles include privacy, accountability, safety and security, transparency and explainability, fairness and nondiscrimination, human control of technology, professional responsibility, and promotion of human values.

But how do these principles get applied? Everyone supports "fairness" in the abstract, but people disagree vehemently about what is fair in practice; just look at the evolution of debates around affirmative action in hiring and college admissions, or reparations for the descendants of enslaved people. Likewise, the aspiration of "promoting human values" means almost nothing on its face: Liberty and equality are both broadly held human values, for example, but the pursuit of one can come at the expense of the other.

The development of responsible, ethical, and beneficial AI is complicated by the reality that different people have different definitions of "responsibility," "ethics," and "benefits." What benefits one group often harms another—managers and workers, competing firms, urban and rural residents, rich and poor. In these contexts, agreement about AI ethics can only be reached on the tamest of terms. The signatories to the nonprofit Center for AI Safety's "Statement on AI Risk"—a hundred prominent computer scientists, CEOs of AI companies, Google alumni, and other luminaries—agreed on a single sentence: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." Who could disagree with that? But how does such a statement reduce these hyperbolic risks, never mind make a better chatbot?

This emerging technology requires more than abstract lists of ethical principles or positions that connect at the lowest common denominator. As policy scholars and moral philosophers cast around for consensus principles, they should explore the fact that every society already operates on a set of near-consensus beliefs—what policy scholar Barry Bozeman calls *public values*. Public values are the common sense of a political community: the goals that almost everyone endorses, even if they disagree on how to achieve them. Building better mechanisms for engaging public values in the development of AI policy could be a more productive way to gauge and pursue shared conceptions of desirable futures.

Where there's agreement

Public values, according to Bozeman, are "those values providing normative consensus about (a) the rights, benefits, and prerogatives to which all citizens should (and should not) be entitled; (b) the obligations of citizens to society, the state, and one another; and (c) the principles on which governments and policies should be based." Once you start looking, you notice public values everywhere in public life. They are written into constitutions, major pieces of legislation, and court decisions; carved into monuments, and invoked in the

speeches, songs, poems, and stories that make up a political community's cultural canon.

Public values form the fundamental aspirations of society: safe communities, healthy lives, meaningful work, freedom from want, fear, and oppression. In a survey of more than 2,000 US citizens, Bozeman identified an impressive number of consensual public values. Over 90% of respondents supported substantive principles of freedom of speech, physical liberty, equal access to civil rights, freedom of religion, gender equality, and physical safety and security; and over 80% supported the protection of minority interests, access to health care, economic opportunity, and privacy. Many more goals would pass the 90% mark in smaller jurisdictions, particularly individual states. Public values are not eternal, unchanging, or universal. But they tend to be widely recognized and change slowly, emerging as they do from the long sweep of political contestation, negotiation, and compromise.

Public opinion polling plays a crucial but limited role in the understanding and formation of public orientation toward emerging technologies. Opinion polling captures only a snapshot of a policy issue in flux. Citizens' views of emerging technologies may change rapidly as they learn more about them, but the values informing those opinions tends to remain consistent. This limits the applicability of the usual public opinion approach of asking citizens to rate agreement with particular policy proposals. In principle, opinion polls (like Bozeman's survey) could ask about public values, and they should. Yet more crucially, public opinion polling does not, on its own, help citizens to better engage with emerging technology, their fellow citizens, and their points of common ground and disagreement.

Instead, the practice of uncovering public values—particularly ones that characterize individuals' relationships to technology—depends on a combination of detection methods. Decades of research in technology assessment, responsible innovation, and anticipatory governance have yielded many approaches to examining and shaping societal consequences of emerging technologies. These include public forums, scenario planning methods, and integrating social and humanistic researchers into the technical research and development process. Of particular importance are deliberative methods which give citizens space to learn about technologies, discuss them with their peers, and take time to articulate their own priorities and rationales. Such considered positions are less prone to rapid flux. Moreover, deliberative methods give citizens and researchers the opportunity to discover priorities together, and to articulate how today's technological issues connect to more durable conceptions of the good society and the good life. More support for investigating public values via deliberative approaches is urgently needed.

Even though Americans are profoundly polarized, the process of examining deeply held public values can enable the discovery of common ground. According to Pew Research, few Americans support the use of AI to advise people about faith in God, matchmaking, and governing the country. Congress has already prohibited, in the TAKE IT DOWN Act of 2025, dissemination of both real and AI-generated nonconsensual explicit content of individuals, acting against a growing and highly disturbing form of harassment. But for politicians and citizens alike, the potential for political realignment can only be realized by engaging with one another and with the situation at hand.

AI ethics principles can, and hopefully will, eventually reflect more than just the discourse of experts and computer scientists—but only by being debated, hammered out, implemented, and modified in actual communities, governments, and businesses. Decisionmakers can get a head start by building upon public values, which have already emerged from this process of iterative trial, implementation, and deliberation.

Operationalizing public values in AI

Public values are embedded in every law, regulation, and standard. The origins of most modern regulatory agencies, like the US Food and Drug Administration, the Environmental Protection Agency, and National Highway Traffic Safety Administration, can be pinned to a few foundational public values. These agencies were developed and advanced by broadly held notions about public safety and environmental health. They were built on widely shared dismay at the ravages of snake-oil medicines, adulterated foods, polluted air and water, and traffic deaths—ideals founded in even older social traditions of mutual respect and environmental stewardship.

Public values can be expressed as a specific list of widely endorsed, substantive positions about what society should look like, drawn from the shared history and political culture of real nations and communities. Such a list can constitute a map of issues to consider in the design, implementation, and regulation of novel technologies, as well as to assess how technologies are affecting the things that people care about.

For example, in our recent book, *Advanced Introduction to Innovation and Public Values*, Bozeman and I articulate a heuristic framework for assessing the societal implications of emerging technologies, built around the following list of public values: civil liberties, equal rights, democratic governance, rule of law, national security, public safety, widely shared economic prosperity, public health, and environmental health. People do disagree about what these values mean in particular contexts—especially pertaining to novel technologies—but as narratives are reshaped by new ideas and experiences, areas of agreement can be discovered.

What would it mean for AI design, governance, and implementation to be guided by public values? Policymakers, scholars, technologists, managers, and advocates can use a public values framework to identify societal risks and opportunities emerging from AI. They can track how AI firms and AI policies are impacting public values, which can be used to guide possible approaches for federal and state policy. The table below includes

examples of a range of policy tools available for shaping AI development and implementation aligned with public values.

Furthermore, researchers and engineers could engage with citizens and social scientists to align technology design and implementation with public values in the process of development. Firms' standards of social responsibility, in ethics and compliance offices and boardrooms, could be reconceived to take public values into account. Regulators and judges could consider public values as substantive elements of public interest in deciding how to shape the law and resolve disputes. These shifts could be supported by new infrastructure for engaging citizens as they decide their views on AI and how they would like it to fit into their futures, such as formal and informal opportunities to learn and deliberate across political communities.

For past technology revolutions such as mass agriculture, pharmaceuticals, automobiles, and plastics, the tools for shaping the effects of technologies on society (regulation, standards, education, labor organizing, consumer advocacy, etc.) were deployed only after severe negative consequences emerged. In the case of AI, they can and should be deployed in a more agile and anticipatory way, to proactively advance public values.

POLICY IDEA	RELEVANT PUBLIC VALUES
A requirement that all "deepfakes" depicting real persons, events, or issues of public significance be conspicuously labeled as such, with criminal penalties for disseminating unlabeled deepfakes.	Public safety, democratic governance
Criminal liability for providers of large language model chatbots which are found to provide aid, direction, or encouragement to persons in committing violent crimes or suicide.	Public safety
A requirement for conspicuous notification for users or customers about whom certain important decisions (e.g., loans, insurance coverage) will be made by automated processes, and a right of easy and timely appeal to a human review and decision process.	Due process, equal rights
A requirement that major data center construction projects develop community benefit plans addressing, among others, labor and environmental concerns.	Widely shared economic prosperity, environmental health
A requirement for substantial community and employee representation on the boards of AI firms operating in a given state.	Public safety, widely shared economic prosperity, environmental health
Construction of large-scale computational resources at state universities, ensuring that cutting-edge computation is available for publicly oriented science.	Public safety, national security, widely shared economic prosperity
Workforce training and education to upskill vulnerable workers in emerging technologies and to prepare citizens for the dangers of AI in political discourse.	Public safety, widely shared economic prosperity, democratic governance

Laboratories of AI governance

It is in the work of politics, negotiation, persuasion, and deliberation that public values actually develop, emerge, and become durable. Agreements are built through political interchange. Democracy and governance, no less than science and engineering, are experimental arts. Experts or leaders cannot project *a priori* what goals citizens will have, nor what governance arrangements will best protect and advance these goals; these develop through a continuous and dynamic process of trial, error, and adaptation.

In the United States, states are "laboratories of democracy"—jurisdictions that can undertake and learn from divergent policy experiments. States can and should experiment with regulation, standard-setting, public-private partnerships, public goals–conditioned funding, and other efforts to protect their citizens from AI's downsides and ensure they capture its upsides. Several states have already exhibited leadership in AI regulation informed by public values. Colorado has adopted a comprehensive limitation on high-risk AI systems similar to the European Union's Artificial Intelligence Act. Texas prohibits the development of AI systems for discriminatory purposes.

In the spirit of learning how to govern in the age of AI, states should be proactively and fearlessly looking for ways to protect, support, and advance citizens' values by shaping AI technology, infrastructure, and markets within their jurisdictions. Not all these experiments will work. But in governance, as in technology development and in business, learning requires a willingness to experiment and fail. It is only in the negotiation, fray, and experimentation of politics that values are settled on and the merit of different ways to advance them is adjudicated.

Democratic governments are the core institutions by which societies manifest public values. It is imperative that these institutions not tie their own hands when it comes to important new issues like AI, as the recent bid to prohibit US state regulation of AI would have done. Now is not the time for governments to abdicate responsibility for ensuring that a potentially world-changing new technology genuinely benefits most people.

The European Union and China have already adopted policies to protect their citizens from obvious physical, social, and privacy harms caused by new digital technologies such as AI. The United States' federal, state, and local governments must not fall too far behind. Moreover, governments and firms, along with engineers, researchers, and advocates, should do more than merely prevent AI from causing direct and obvious harms. Rooting future aspirations in deeply held public values holds promise for building a future in which most people actually want to live.

John P. Nelson *is an assistant professor in the School of Public Policy at Oregon State University.*