Working Paper

How Can AI Help People Become Better Versions of Themselves?

Benjamin Boudreaux, Robert J. Lempert

RAND Pardee Center

WR-A2717-1
August 2023

For more information on this publication, visit **www.rand.org/t/WRA2717-1**.

About RAND

The RAND Corporation is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest. To learn more about RAND, visit www.rand.org.

Research Integrity

Our mission to help improve policy and decisionmaking through research and analysis is enabled through our core values of quality and objectivity and our unwavering commitment to the highest level of integrity and ethical behavior. To help ensure our research and analysis are rigorous, objective, and nonpartisan, we subject our research publications to a robust and exacting quality-assurance process; avoid both the appearance and reality of financial and other conflicts of interest through staff training, project screening, and a policy of mandatory disclosure; and pursue transparency in our research engagements through our commitment to the open publication of our research findings and recommendations, disclosure of the source of funding of published research, and policies to ensure intellectual independence. For more information, visit www.rand.org/about/research-integrity.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Published by the RAND Corporation, Santa Monica, Calif.

Limited Print and Electronic Distribution Rights

About This Working Paper

Artificial intelligence (AI) augurs changes in society at least as large as those of the industrial revolution. Much current policy debate seems more narrowly focused on extrapolating current trends and asking how to manage the risks, ranging from distracted attention spans to human extinction. But AI could enable radically different futures in terms of how people live and work, the values and choices people can pursue, and risks they will face. This essay offers an initial exploration of the near-term policy implications of one aspect of this potential for a transformed, AI-enabled future. The essay begins with the premise that humanity has already demonstrated the ability to develop algorithms that change people's personalities and values. To date many of these changes are side effects of private sector firms seeking to generate revenues. But what if algorithms and AI more broadly were purposely designed to change people's personalities and values for the better? With such capabilities, AI might truly enhance human well-being. But this possibility raises thorny question for a liberal society: Who (or what) would get to decide what it means for people to be better? Who would build and operate the algorithms? What roles should government, business, and civil society play in governing and influencing such algorithms and the ways in which people interact with them?

This Working Paper aims to engage an initial discussion on such questions, which are not only of intrinsic interest, but also one gateway into a wider discussion regarding how best to shape the future into which we may be heading. This work was funded by the RAND Pardee Center for Longer Range Global Policy and the Future Human Condition.

About the RAND Frederick S. Pardee Center for Longer Range Global Policy and the Future Human Condition

The RAND Frederick S. Pardee Center aims to enhance the overall future quality and condition of human life by aggressively disseminating and applying new methods for long-term policy analysis in a wide variety of policy areas in which they are needed most. There has been no shortage of past attempts to think globally about the human condition or the long-range future. What has been missing, however, is a means of tying those efforts systematically and analytically to today's policy decisions. This is the gap the Pardee Center seeks to address.

Questions or comments about this essay should be sent to its authors, Benjamin Boudreaux <bboudrea@rand.org> and Robert J. Lempert (Lempert@rand.org). Information about the Pardee Center itself and its other projects and initiatives is available online (www.rand.org/pardee). Further inquiries about Pardee Center activities and projects should be sent to the Pardee Center Director, Robert J. Lempert, at Lempert@rand.org.

Summary

Humanity has demonstrated the ability to develop algorithms which can attract people's attention in ways that change their personalities and values. To date, many of these changes have been undesirable side effects of algorithms designed to sell advertising. But what if algorithms and AI more broadly were purposely designed to change people's personalities and values for the better? Such capabilities might enable AI to truly enhance human well-being. But the possibility also raises thorny question for a liberal society: What it would mean to change people's values, desires, dispositions, or even personalities for the better? Who (or what) would get to decide what it means for people to be better? Who would build and operate the algorithms? What are the roles for government, business, and civil society in governing and influencing such algorithms and the ways in which people interact with them?

In attempting to address such questions, this piece purposefully tries on an optimistic view, with the intent of exploring the conditions under which these ubiquitous algorithms might contribute to human well-being. We begin by reviewing the institutions and social structures -- courts, schools, churches, families, etc.—that currently shape values in liberal societies. We then review three ethical views about the good life that might guide how AI could be used. Of the three -- – hedonic, utilitarian, and an approach based on capabilities and deliberative dialogue – we suggest that only the third is appropriate.

We then suggest the implications of this capabilities and deliberative dialogue approach for near-term policy action. Given that AI presents a society-transforming technology that will almost certainly change people's preferences, values, personalities, relationships, and worldviews, we explore what actions might be taken today by whom, the processes by which groups of people might determine what constitutes "better;" and the implications for how AI ought to be designed. We offer some initial speculations on these questions, a discussion of algorithms designed to facilitate human deliberation and decision-making, and conclude by offering with principles that might guide the evaluation of specific policy proposals. These principles are: centering human well-being, co-production, equitable access to AI as a public good, and transparency for AI while maintaining privacy for humans.

Introduction

Humanity has demonstrated the ability to develop algorithms which can attract people's attention in ways that change their personalities and values. To date, many of these changes have been undesirable side effects of algorithms designed to sell advertising. For instance, social media algorithms are designed to make the people interacting with them more predictable (and thus more likely to click on certain content) and as a side effect make them angrier, more extreme, and more polarized.[1] There is also increased evidence of a deleterious effect of social media on youth mental health, resulting in a recent Surgeon General advisory.[2]

But what if algorithms and AI more broadly were purposely designed to change people's preferences and values for the better? Such a question builds on those raised by the alignment problem (Christian 2021), which aims to steer AI systems towards betterer serving human goals and interests. Discussions of the alignment problem often assume human values are fixed or not appropriate to change or challenge. But this is clearly an extremely limiting assumption. New technologies with the potential significance of AI have always had profound effects on human values and relationships. The printing press helped launched the Reformation, re-orienting many Europeans' views on the relationship between humans and God. The large-scale changes in technologies and economies of the industrial revolution were inextricably embedded with 19th and 20th century changes in political, religious and social relationships (Polanyi 1957). As one example, inexpensive transportation and the accompanying market revolution changed American concepts of democracy, family, and religion at the start of the United States' industrialization (Sellers 1994). The science, technology, and society literature highlights such intertwined changes in technology, technology systems, and human values (Köhler et al. 2019).

Treating peoples' values as exogenous makes it difficult to consider perhaps the most fundamental questions involving AI – how can the technology improve individual and societal well-being, including providing opportunities for human flourishing and enabling more just social institutions? Answering such a question may well involve designing human interactions with machines such that people's values, as well as the objectives of the algorithms, evolve towards something better than they were.

---

[1] See, for instance, https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/, https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html, https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html, https://www.cjr.org/the_media_today/youtube-radicalization.php, and https://bhr.stern.nyu.edu/polarization-report-page

[2] https://www.hhs.gov/sites/default/files/sg-youth-mental-health-social-media-advisory.pdf

The possibility raises thorny questions for a liberal society dedicated to pluralism and diversity: What it would mean to change people's values, desires, dispositions, or even personalities for the better? Who (or what) would get to decide what it means for people to be better? Who would build and operate the algorithms? What roles should government, business, and civil society play in governing and influencing such algorithms and the ways in which people interact with them?

In attempting to address such questions, this piece purposefully tries on an optimistic view, with the intent of exploring some of the ethical challenges of using these ubiquitous algorithms to enable a much better world. Positive, hopeful visions are useful to consider because they can provide motivation for action. Liberal societies currently contain many institutions and other societal structures – courts, schools, churches, families, etc. –designed to channel human behaviors in directions deemed desirable and away from behaviors seen as undesirable. While these institutions have provided benefits, they often operate far from their ideal, failing to address patriarchy, racism, violence, poverty, inequality, and environmental destruction. This essay explores a best-case scenario for how liberal societies might use algorithms to overcome such challenges and help people to live better lives. We recognize that there are many ways in which the scenario presented here might fall short. But our purpose is to explore how algorithms might be used in the best cases, not as a prediction that this will come about, but as an attempt to articulate an ideal that might guide policy and debate. Subsequent work would turn to the necessary questions of whether this ideal state is possible, and of how some of the ideas explored here might go wrong.

This piece focuses on algorithmic recommendation systems associated with social network applications that offer up content designed to engage users. These systems include those within social media platforms such as YouTube, Facebook, TikTok, and Twitter. While our discussion will have implications for emerging generative AI systems such as large language models that statistically predict which words follow others, these generative AI systems are not a principal focus of this piece.

How Liberal Societies Currently Shape Values

Psychologists generally distinguish between preferences and values. The former are individual inclinations to favor one option over another, such as preferring a particular color for a sweater or having a favorite movie. Preferences vary among individuals and may depend strongly on context. Values represent more enduring and deeply held core beliefs regarding what is important or good in life. Values guide a person's behavior over time and undergird human

identities, relationships, and the goals they pursue. Nonetheless, values also respond to context since people's life circumstances provide opportunities to pursue or express some values more easily than others (Schwartz 2006). For instance, having economic security and a job that incentives creativity may enhance the salience of values that favor self-expression relative to values favoring security and conformity.

Liberal societies primarily regard adults' values as their own business. Some constraints and caveats related to this claim exist, for instance citizens are expected to support democratic processes and tolerate those with whom they disagree. State laws enforced by punitive measures shape citizens' actions, but these laws are intended to leave broad room for citizens to develop their own distinct comprehensive conception of what is good (Rawls 1971).

Liberal societies do however contain many organizations and activities aimed at shaping peoples' values. These include voluntary organizations such as churches, clubs, and other civil society groups. Communications such as advertising by businesses, the media, art, and political discourse are primarily focused on influencing behavior, but also shape values. Thus, some organizations and activities in liberal societies explicitly aim to help people become better versions of themselves, while others aim to shape values for other purposes. In principle though not in reality, adults can choose which of these organizations they wish to associate with and what messages they listen to. In practice, some organizations and messages may be difficult to avoid and others difficult to find or difficult to find acceptance within. In addition, individuals' choices are strongly shaped by their social networks. In general, the government is constrained to shape behaviors through law and incentives, though shaping behavior can shape values regarding appropriate behavior. The government is also allowed to influence people's beliefs on issues regarding public health and safety, for instance sponsoring advertising aimed at delegitimizing smoking and drunk driving or promoting the use of seat belts or regular exercise and good diets. Political conflicts do arise, however, when laws (say around public health) are seen by some citizens and associations as infringing on personal liberties.

Figure 1 sketches an admittedly simplified and ahistorical web of societal influences on adult individual's values, including friends, family, civil society such as religious or other groups, the economy often through media and advertisements, public discourse, and the state. Note that liberal societies place some rules on how economic entities and voluntary organizations (including churches) can behave. While generally promoting free speech, liberal societies have rules and norms that limit some types of speech to prevent physical or reputational injury, conform with some societal norms, and protect the marketplace of ideas from false information. Even the U.S., with relatively aggressive protections of unlimited speech, prohibits knowingly false statements in commercial advertising, libelous speech, and some types of obscenity. In general, political discourse is less restricted than commercial discourse.

With children, liberal societies explicitly recognize the importance of shaping values and craft networks of individuals and organizations designed to do so in desired ways. Such individuals and organizations include families, churches, schools, and organized recreation (e.g., sports, etc.). The family can play a more significant role for children than adults in directly influencing the most deeply held values and the other people and institutions with which they interact. In addition, children's values are more readily shaped and shifted than adults. The forces that operate on children therefore have a greater influence and thus society recognizes these entities' special responsibilities.

Of course, this is not a neat picture—contextual and historical factors significantly shape this dynamic for both children and adults. All societies favor and facilitate material advantages to certain positions over others, and the social forces depicted in the figure will similarly reward a certain set of values and positions. For example, some professional or other positions are regarded 'high status', some relationships and familial structures are considered normative while others are considered deviant, standards of aesthetics and other norms suggest who is beautiful and who is not, who or what is considered friendly or threatening, and so forth. These historical and contextual factors operate over time to give rise to accumulated advantages for some and accumulated disadvantages for others. In brief, not all persons are similarly situated in the system. Citizens' values are shaped in the context of this inequity, where the economic/social opportunities, and how one feels about oneself, one's body, and the broader community is dictated by often unjust circumstances associated with the natural lottery. Although the different associations and groups might have continued influence over the entire lives of persons, there is also a process of learned habituation that tends to 'lock-in' certain beliefs and dispositions, which thereby leads these persons to seek out associations and communities that align with those beliefs and dispositions.

How algorithms currently affect people's values: Commercial advertising exists to shape peoples' preferences and there is some evidence that it works (Defever et al. 2011). Operating under the current norms and rules, today's tech firms have developed algorithms designed to predictably capture people's attention by tailoring and feeding them content; the content that tends to most capture people's attention tends to be content that makes them anxious, angry, and polarized. It is currently unclear the extent to which interacting with AI algorithms causes people to express some values more than others (Guess et al. 2023). Our premise is that the possibility that algorithms could have that effect, now or in the future, is worth considering.

Figure 1 thus includes the relatively new addition of social media algorithms, largely shaped by economic actors, which may now play a significant and often corrosive role in shaping the values of individuals and the other people and institutions with which they interact. The large technology developers claim that their products improve some aspects of people's lives—for instance connecting the world and making information widely accessible—but acknowledge only a small part of what we now know are the many unintended consequences associated with

people's dependence on and prodigious use of their products. Even if individuals within these companies seek to act ethically, the companies themselves operate within an economic system which incentivizes them to build and deploy tools that do not support peoples' ability to critically deliberate about values and sense of the good life.

How AI Might Promote the Good Life

It is a bit of an understatement to suggest that the current situation is inadequate. Society is unleashing new technology that is and will continue to change how people behave, what they believe, and what they value within a socio-technical system that serves only a narrow set of interests. In many respects, existing social media technology has already influenced people's values and interactions for the worse, and emerging technology such as AI chatbots stands to further exacerbate this dynamic. It thus seems important to attempt to envision a world in which people can choose to work with and leverage algorithms designed with their broader interests in mind. The idea is that rather than companies treating users as a means to captivate attention and extract data, perhaps we can describe a situation where people use technology in ways that serve their considered interests. In such a world the social network algorithms might have significant influence, but would themselves be shaped largely by individuals, their friends and associates, and civil society elements shown in Figure 1.

The question then becomes, how might we envision a world in which AI and in particular, social network algorithms, help people to interact with both each other and the algorithms in ways that make all the individuals involved better versions of themselves.

For some of the substantive values that should both guide the deployment of AI and that AI might support (given appropriate governance environment), we first look to the values that provide the social foundations of just liberal democratic states. These include values such as familiar bedrock human rights and civil liberties, including privacy, non-discrimination and equity, accountability and transparency in government, and the like. These shared political values undergird the solidarity and social unity of the community while also providing a structure for the exploration of more personal values and practices such as in the context of interpersonal relationships, spiritual contexts, the pursuit of knowledge and the like. We do not prescribe these personal values, but instead emphasize that they come from the critical reflection of individuals' engaging in social environments and with the support of just political institutions.

To explore how AI might help its users better realize these political and personal values, we first need to consider some conceptions of what philosophers (and others) have often called the good life. This question goes back through human history and finds diverse answers in

perhaps all religions and cultures, and so obviously we cannot review the myriad of possible answers. But we can review two views often associated with discussions surrounding AI, before turning to an alternative view that we think provides an appropriate foundation for enabling a liberal society to use and govern an AI that helps people become better versions of themselves. This discussion draws on Parfit's discussion "What Makes Someone's Life Go Best" (Parfit 1984).

Hedonic view: The first type of answer holds that the good life consists of pleasurable mental states. These pleasures might be from carnal experiences of a delicious meal or the touch of skin, or "headier" experiences associated with interesting conversation or watching a good movie.

We need not dwell on the specifics of the view, except to note that this sort of good life might be one that algorithmic systems and AI more generally is particularly good at helping enable. Indeed, technologies such as the metaverse or even video games provide some users with a significant amount of pleasure, and as AI learns more about us and becomes more powerful, these technologies may help us experience more compelling forms.

However, the pleasure-oriented view runs into some significant ethical objections. On this view, we could be living a completely wonderful life while being plugged full-time into an experience machine that simulates things that give us pleasure (as was highly dramatized in the Matrix). This runs contrary to a strong intuition that a good life must consist in part in having real relationships and doing real things rather than having pleasurable mental states created by simulations or manipulations. The fact that AI might be so good at promoting pleasurable mental states to the detriment of real experiences is a problem for AI, rather than something that should be celebrated. The hedonic view provides little guidance on how to guard against such dangers.

Utilitarian view: A second conception of the good life is not about pleasurable mental states per se, but about preference or desire satisfaction. Economists often adopt this view in modeling human behavior with a focus on what psychologists regard as preferences, not values (Warren et al. 2011). Preferences define utility functions that rational actors seek to maximize subject to constraints. Many leading thinkers on the AI alignment problem also adopt this view, basing their work on a utilitarian understanding of ethics focused on achieving outcomes assessed as good based on utility optimization (e.g. Russell 2019).

A person might achieve pleasurable mental states when a preference (for instance for a tasty meal) is met, but there also exist many other preferences that do not produce pleasurable mental states when satisfied. For instance, a person might reasonably have preferences to sacrifice pleasure for the sake of some other objective, or preferences that other individuals achieve something even if the preference holder never knows about the achievement. Further, many of our preferences require that we actually do something. Merely simulating that we've won a prize or helped out a friend would not actually satisfy such preferences.

In these ways the preference satisfaction view is distinct from the pleasure view, and thus the main objections to the pleasure view do not hold. But the utilitarian view also faces significant objections, namely that not all preferences are valid in enabling a good life. There are two broad critiques that underscore a refined approach.

First, some preferences aren't worthy of being satisfied. We would regard unworthy of being satisfied preferences based on false beliefs, those shaped by historical trauma or overly restrictive social norms, or that involve the subjugation or abuse of others. To the extent that a person does hold such preferences, their satisfaction would not contribute to what others and perhaps themselves might regard as a good life.

Second, preferences may be inconsistent with each other and insufficiently considered to support what an individual might upon reflection regard as a good life. Many people and philosophic traditions emphasize the importance of reflection and introspection over preferences and the importance of learning over time. Such reflection and learning aims to enrich our values as we come to better understand the implications of acting on our values, as we attempt to explain our values to others, and as we learn about others' values. Such processes are inherently social and takes place in the context of communities and politically structured institutions. Similarly, we should understand preferences not as atomistic elements possessed by detached individuals, but as shaped in communal contexts. While communal interactions can certainly foster exclusion and group-think, regular social interactions are often necessary to ensure values such as fairness and respect for diversity within a shared common humanity.

While current algorithms excel at satisfying people's actual and sometimes ill-considered preferences, they are not well-crafted to help people reflect on their preferences and consider what values might help them achieve a better life. Indeed, the rapid evolution of algorithms over the past decade is based on machine learning models that learn from us, including our biases and often insufficient inclination to reflect on our choices. In this way, algorithms designed to learn from our existing preferences can reenforce our worst impulses rather than creating opportunities for us to adopt better preferences over time.

Capabilities and deliberative dialogue approach: Neither hedonism nor utilitarianism provides an appropriate foundation for considering how a liberal society can use AI to help people become better versions of themselves. The importance of additional normative considerations beyond current preferences, along with learning as a means to discover such considerations, suggests a third approach to the good life that might provide that foundation.

This view of the good life, which we call the capabilities and deliberative dialogue approach, involves three key elements (Sen 2001, Sen 2009) McCoy and Scully 2002). First, there exists some objective list of elements that in some combination can be used to evaluate both individual lives and societal choices independent of any particular individual's current preferences. This objective list might include things like meaningful human connection, intellectual activity,

spiritual practice, possessing moral virtue, or acting in accordance with the moral law. Second, this view recognizes inevitable diversity among individuals and cultures on the specific elements on this objective list and the priorities among them. Many would regard this diversity as not only a fact but a treasured attribute of the human condition. Third, this view highlights the importance of how choices are made. A good choice is one that is freely made, duly considered, and based on reasoning that one can explain to oneself and others. Sen (2009) compares choosing a quiet evening at home to house arrest. While the physical circumstance are similar, the former is highly preferable because it results from a freely made, considered choice.

For the objective list of elements needed to consider any betterment of the human condition we adopt the "capabilities approach," a well-known theory articulated most prominently in differing forms by Martha Nussbaum and Amartya Sen. Sen (2009) begins his inquiries into human betterment by arguing that the goal is a society in which people have the capacity to live in ways that they find reason to value. Potential improvements to society ought to be judged according to the extent to which they increase such capacity. Individuals will of course differ in the lives they wish to lead, but Nussbaum and Sen argue that the variations can be encompassed by a list of capabilities. These capabilities include bodily health; bodily integrity; sense, imagination, and thought; emotions; practical reason; affiliation; other species; play; and control over one' environment. Activities that people have reason to value are not purely a function of their current preferences but also include consideration of the kind of beings that humans are and the sort of opportunities that help them flourish. In contrast to preference-satisfaction theories of well-being the capabilities approach provides a framework for considered reflection by individuals and groups that critiques and adjusts preferences to better align them with living lives that people have reasons to value.

The capabilities approach includes an assessment of what makes someone's life go well, but also guides consideration of the political institutions and policies that promote just social arrangements, that is, equitably create capacities for people to live in ways that they value. Social institutions that help support these capabilities include mechanisms that encourage freedom from coercion by others and dismantle existing systems of oppression as well as publicly supported resources such as education and healthcare and socio-political conditions that allow for capabilities including political participation. Although resources are important to exercise capabilities and current conditions of inequality are a major hindrance, the capabilities approach holds that well-being is not rooted in any material conditions, but instead consists in the opportunities to conduct valuable activities—in other words, social institutions should be assessed by a standard of what people are able to do rather than what people possess.

Table 1 adopts a list of capabilities suggested by Nussbaum to explore the extent to which current algorithms contribute to various human capabilities. For many of these, algorithms might be able to support capabilities, but only in contexts where the algorithms are supplemented with extensive real-world conditions and activities, and are governed effectively. For others, current

algorithms actively distract and derail capabilities, such as opportunities for meaningful connections or bodily health.

This table helps suggest ways in which current algorithms are not at all well-designed to support humans in reflecting individually and with each other on what it means to have a good life. As is discussed further below, one major hurdle is the economic model in which algorithms are developed—large tech companies with access to significant compute and data that deploy them to capture people's attention—rather than designed and deployed intentionally to support human well-being. Thus, we believe an alternative approach that shifts this economic model is required.

How, in a world of diverse visions of the good, should society make choices as to which policies and institutions best foster the capabilities that individuals require to lead lives that they value? Traditional social choice theory emphasizes markets and voting as primary means of solving collective action challenges. Sen and other advocates of capabilities theory also emphasize deliberation as an important complement to the latter. This focus is important because it can suggest ways in which algorithms can enhance human well-being by enhancing effective deliberation among humans.

Deliberative dialogue involves the way that members of a community express their views and build collective approaches to shared problems. Such deliberation can help people critically reflect on their preferences (and blind spots) and in some cases shape these preferences in ways that are more aligned with enhancing individual and community-wide capabilities for pursuing lives people value. Deliberation can also enhance self-knowledge so that our preferences become more authentic, reflective, and considered. Deliberations work best when they recognize the inescapable plurality of competing views; facilitate re-examination and iterative assessments; demand clear explication of reasoning and logic; seek consensus on near-term actions that can be taken rather than on what constitutes ideal institutions or an ideal good life; and recognize an "open impartiality" that accepts the legitimacy and importance of the views of others, both inside and outside the community of interest to the immediate policy discussion.

Implications for Action

We can now return with some initial answers to our framing question. AI presents a society-transforming technology that will almost certainly change people's preferences, values, personalities, relationships, and worldviews. If actions were to be taken today with the aim of ensuring those changes were for the better, who or what should get to decide what constitutes "better" and the implications for how AI ought to be designed?

How algorithms might enhance decentralized human decision making: Our answer is that AI, in particular the algorithms that seek to facilitate social connection and make recommendations about what sort of content or people with whom we ought to engage, along with the socio-technical system in which they are embedded, ought to emphasize decentralized and community-based decision making; a diversity of visions of the good; rich communication among groups holding different visions of the good; ongoing iteration that strives to move all views closer to their ideals; and non-domination (Allen 2020). The goal is to ensure that no small set of actors dominates decision making on algorithm design or its socio-technical systems and no visions held by any group unduly imposes itself on individuals. Instead, algorithms should be used to help individuals and groups identify and pursue their visions of a meaningful life, free of inappropriate economic or social influence by others but consistent with obligations to the community. This includes helping to identify and dismantle existing systems of domination that unfairly enable the benefits of society to accumulate to some races, genders, and cultures to the disadvantage of other as well as a just distribution of materials resources consistent with planetary environmental health. These are aspirational, hard to achieve goals for a good society independent of the influence of algorithms. The hopeful vision is that algorithms can be configured to help society move closer to these goals.

Clearly, this is not the way in which such algorithms are currently developed. Rather, these algorithms emerge from an economic model in which profit-seeking firms shape the technology for their own benefit in a largely laissez-faire governance context. The developers and implementers of AI have their own economic interests and operate in a regulatory environment that aligns them insufficiently well with the public interest. Public governance has been largely lacking and otherwise slow for several reasons. There is a widespread ideology among political and economic elites that seeks to support technological innovation for its own sake, which contributes to fears that government regulation will slow or stall AI development. The competitive dynamic among large for-profit technology companies to be the first to market is further exacerbated by rhetoric that US and China are in an international race for AI. The current governance environment thus cedes judgements to large technology companies on what technologies and business models best advance human betterment, with no mandated Federal standards for transparency that would allow greater public accountability, much less forums for public deliberation and guidance on these important decisions.[3]

Along what pathways might such guidance send AI technology and the socio-economic systems in which it is embedded? Such an AI might be designed to help enable a society with largely polycentric governance among many independent yet interdependent agents – individuals

[3] Consider for instance the White House announcing recent voluntary non-binding commitments made behind closed doors by major AI companies (see https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/)

and groups – pursuing their goals with mixes of competition and collaboration.[4] Each such agent might have an AI customized to help them achieve their goals, though subject to a set of baseline standards and safeguards negotiated and implemented by relevant governance structures. The actions and associations of these agents help individuals to explore and express their multiple individual and group identities. These identities span and interact with each other and across spatial scales, including individual, communities, national, and planetary. At the level of individuals and small groups the vision of the good life/human flourishing may be relatively specific, organized around specific activities and practices. In larger groups, the vision of the good life may be more general meta-concepts, plastic enough to adapt to local needs and practices yet robust enough to maintain a common identify and capacity for collective action across groups (Ansell 2011). In larger groups, communities will also need to determine how to work collaboratively on shared challenges—say environmental sustainability—and other contributions to the world around us.

Alternative models of algorithm design: These comments address societal level institutions, norms, and policies. But these processes also have implications for the design of AI systems themselves. Currently, AI is designed as autonomous decision-making agents pursuing well-defined objectives, but that tailor their outputs based on individualized learning. Some proposals in AI alignment aim to ensure that these objectives serve human preferences and to create AI that recognizes it may not understand those preferences so seeks to learn more about them. But at its foundation, current algorithm design focuses on what the decision support literature would call the choice task: choosing among a defined set of options to find the one that takes you closer to your objectives (Gong et al. 2017).

Decision support analytics can, however, perform a distinctly different task called decision structuring (Helgeson 2018), which includes defining the problem in a way that opens it up to thoughtful consideration, defining the objectives to be achieved, and assembling a menu of options that might achieve those objectives (Parker and Fischhoff 2005, Del Missier et al. 2014). Much of human decision making, particularly for complex societal problems, consists of iterative processes of framing and reframing, that is, decision structuring, as a prelude to choice (Schoen and Rein 1994). An AI focused on decision structuring would serve less as an autonomous agent and more as partner to individual humans and to groups of humans, helping them to reflect, discuss, learn, and make better decisions. Though they are not a focus of this piece, LLMs in particular might be designed to provide decision structure in ways that can reflect an identified problem and be iterative for an individual and across a community.

In such a role, AI would help humans consider their objectives and the tradeoffs among them from multiple points of view, assemble sets of options, and track progress towards goals. Such an AI might also suggest networks of people and ideas that provide appropriate balances

[4] See, for instance, the Collective Intelligence Project whitepaper: https://cip.org (accessed July 9, 2023).

among re-enforcing chosen actions and questioning those actions from new points of view. Some of these tasks are related to current AI activities. Current Amazon algorithms certainly provide humans with customized options for things they might like to purchase and Facebook algorithms suggests groups individuals might like to join. But at heart, such algorithms are designed as autonomous agents divorced from scrutiny and higher-level feedback from most of the humans they nominally serve. They are not implemented in a way to help a human make a thoughtful decision among a range of options that each have attendant trade-offs, but instead are intended to optimize for content that the human will most likely click without deep reflection.

In contrast, AI algorithms might draw from the design of decision support analytics in the field of decision making under deep uncertainty (DMDU) (Marchau et al. 2019). Decision support represents organized efforts to produce, disseminate, and facilitate the use of data and information in order to improve the quality and efficacy of human decisions. Deep uncertainty exists when parties to a decision do not know or agree upon the probability distribution across future states of the world, the system model that connects actions to consequences, and the weightings of different objectives (Lempert et al. 2003). DMDU analytics are designed, not to identify the best choice, but rather to support analytic-deliberative processes focused on decision structuring tasks. In such processes people deliberate on problem framings, analytics provides products based on those framings, and the humans then adjust their framings in response to the analytic products and deliberations with each other in an on-going iterative cycle (NRC 2009). Such analytic-deliberative processes emphasize learning and work best when participants' expectations and values are expected to evolve over time.

The DMDU algorithms employed in such analytic-deliberative processes perform tasks in support of decision structuring. Such algorithms may search over large databases of simulation model runs in order to identify contrasting scenarios particularly relevant to the decision framings under discussion by the users (Bryant and Lempert 2010), may seek to ensure such scenarios reflect a diverse set of views (Carlsen et al. 2016, Lempert and Turner 2021, Lempert and Turner 2021), or suggest sets of decision options that balance among the competing interests under discussion (Kasprzyk et al. 2013). The goal in each case is to help human users find simple decision framings which help them to make good choices in the face of complex problems. In addition, recent work in social network analysis has also been used to suggest who ought to participate in such engagement, with the goal of increasing equity in reaching all groups within effected communities (Rahmattalabi et al. 2021, Lempert et al. 2023). The tasks performed by all these algorithms might suggest how AI might be designed to even more powerfully enable human analytic-deliberative processes.

The current socio-technical system, in which most AI research and development is conducted by large, private sector, profit-seeking entities and by academics focused on AI as autonomous, choice-making agents, does not seem well-aligned with this vision of a diverse, polycentric system of humans supported by AI. This is not the place for extensive analysis of the

government, civil society, and business policies that might transform the current socio-technical system in directions more aligned with such human flourishing. But such policies might include large-scale, government-supported research and development to encourage co-production with communities and shift technology trajectories; public ownership of the data used to train AI; market-shaping actions such as anti-trust to reduce network monopolies; dis-incentives for advertising-based business models that encourage click-bait; subsidized access for the less well-off so that subscription-based business models remain equitable; industry standards for AI; transparency and auditability requirements; and appropriate IP laws (Rahwan 2018, Parson et al. 2019).

Guiding principles: We can, however, suggest some principles that might guide the consideration of such policies. These principles are: centering human well-being, co-production, equitable access to AI as a public good, and transparency for AI while maintaining privacy for humans. The European Union has developed an AI strategy and supporting regulations (Kop 2021). The Biden Administration has offered a Blueprint for an AI Bill of Rights (OSTP 2022). Both have important overlaps with the principles offered here. However both are focused primarily on harm reduction rather than guiding the creation of new opportunities. The EU regulations are organized around levels of risk and any efforts to deploy algorithms that shape people's values will fall under the highest risk category. The Biden Administration's proposed AI rights – which focus on safe and effective systems, algorithmic discrimination protection, data privacy, notice and explanation, and an opportunity to op out and engage with humans rather than algorithms – largely emphasize freedoms from harms rather than a vision for expanding human well-being. That said, we acknowledge those efforts and suggest they will need to be supplemented with a robust articulation and development of key overarching principles.

It may seem obvious that AI ought to **enhance human well-being**. But the forces currently driving AI development have human well-being as an accompanying, not primary goal. Current AI development is largely driven by market forces, that is commercial investment aimed at capturing private economic returns, and the internal logic of scientific discovery, instantiated as researchers pursuing reputational benefits within academic fields. Both processes are necessary, but not sufficient, to human betterment. Under the current market and other incentives for AI advancement, developers are not responsible for proactively identifying the full range of harms of their products, nor are they under legal obligation to address those harms. They also are not motivated to develop and implement AIs for the goal of human well-being. As suggested by the human capabilities criteria in Table 1 currently AI development contributes only partially to human flourishing and in other cases actively hinders it.

In recent years, there has been a flurry of interest in centering the pursuit of explicit societal goals as the explicit drivers of technology and economic development, with markets and scientific discovery serving as means not ends (Mazzucato 2018). Such processes clearly present

a different set of governance challenges than those which envision largely unfettered markets and discovery as synonymous with the advance of human well-being. But to ensure that AI helps people become better versions of themselves requires explicitly centering enhanced human well-being as the purpose of the AI endeavor. Centering human flourishing will also underscore the limits to AI, and the necessity for people to reorient their attention to human activities and relationships and away from screens, algorithmic feeds, and technologies. There is only so much that AI can and should do in promoting human betterment, and a more constrained and targeted approach to its development will be important for ensuring the role it does play actually serves human interests.

In a liberal society, the **co-production** of knowledge related to AI, and the co-design of AI technologies, is essential to enhancing human well-being. Co-production is a form of knowledge production in which technical experts work together with other groups as equals, each with their own ways of viewing and analyzing the world, in order to generate new knowledge and technologies (Jasanoff 2004). In particular, this process needs to include stakeholders and parties affected by the technology as regularized participants and collaborators in the design and implementation process—especially those that face contextual or social barriers for the inclusion of their considered interests. Co-design is a creative participatory process of fashioning solutions to public policy challenges in which community members and technologists co-produce the necessary technologies and systems (Blomkamp 2018). The co-production concept is foundational to the analytic-deliberative processes described above. Co-production is important for aligning technology development with human values; can make knowledge production more effective by engaging with multiple sources of knowledge, including local knowledge and that from experts in other fields; and is a normative good in that it enhances human agency over the form of the technology shaping their lives. Co-production helps to address the path-dependence of technology development. If a technology evolves for too long without community involvement, it can take a form so that all the options are bad and that specific communities are routinely, even if unintentionally, harmed in ways that lead to accumulated disadvantage over time. Co-production aims to steer technology pathways towards providing people with a better set of options.

To encourage well-being enhancing innovation, a liberal society should employ **democratic shaping of markets**. Liberal societies employ two main approaches for making societal choices: markets and voting (Arrow 1953). Governments in liberal societies, influenced by the votes of their citizens, create the rules which create and shape markets. These markets then create incentives for private actors to change society through their innovation, investment, and consumption decisions. At present, the market incentivizes some constellations of private choices, for instance those supporting business models based on advertising-based social media, that do not entirely align with enhancing human well-being. Co-production can play an important role through all stages of the innovation process, from research design to providing a social

license to operate for individual firms. Nonetheless, public policy through action by government and civil society must also play a crucial role in setting the rules and standards to create market incentives for private actors that better align with human well-being. There of course exists a danger that the choices of democratic institutions, as expressed through government and civil society, can inappropriately slow innovation thereby entrenching powerful incumbent interests and a familiar status quo. Markets have long been seen as a means for breaking down existing hierarchies (Anderson 2017) and this role provided one important justification for the market-based neo-liberal policies over the last decades. However, especially with respect to emerging technology such as AI, the relative balance of power between government and technology companies is out of synch with human well-being. In the current context, technological innovation continues to worsen inequalities of wealth and power. Government and civil society can play an important role in steering markets towards societal goals and using AI to improve governance, thereby reducing the tensions between innovation and democracy might be a valuable focus for AI research.

Enhancing human well-being also requires **equitable access** to AI, to the AI-enabled deliberative processes that pursue human flourishing, and the co-design of the AI that supports these processes. How best to ensure such access is a policy question, but likely requires equity to be included as a fundamental building block in the design of technology and the socio-technical systems in which the technology is enveloped. AI might be useful for identifying barriers to equity and ways in which historical patterns manifest in real-world harms in sectors such as law enforcement, housing, and health care. Co-production in ways that identifies potential impacts across diverse communities will enable those impacted by technologies to be included in design and deployment decisions.

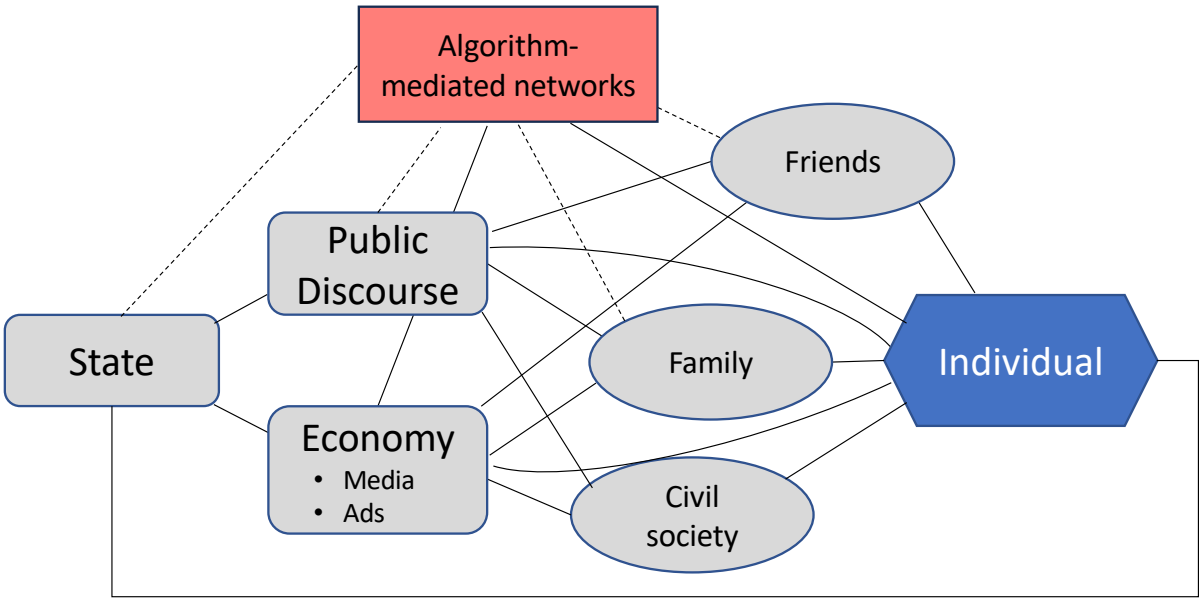
Equitable access and co-production require **transparency for AI**, **with privacy for humans.** Humans require some measure of privacy as a right and a component of well-being. An AI able to help humans pursue their goals requires that the AI have knowledge of individual humans some of which those individuals may wish to keep private. On the other hand, equitable access to an AI co-production process requires that all participants have adequate understanding of the underlying technology and how research choices open and close future pathways. Balancing these needs would seem to require much greater transparency into AI code and the data its uses than is presently the case, along with greater projections for the privacy of individual humans.

Artificial intelligence will likely prove a transformational technology, not only changing the economy and relationships among people but also changing people themselves. Thus, the alignment problem is not just a technical issue in AI design. Rather the challenge is one of aligning policies, institutions, and social systems along with AI so that the entire science, technology, and society system advances human well-being. The present system in which AI development is driven by profit-seeking firms with business models that seek to maximize user

engagement has driven rapid innovation but seems mis-aligned with many dimensions of what many people would regard as a good life. But to align the full science, technology, and society system raises uncomfortable questions for a liberal society – what is a better life and who gets to decide? The utilitarian ethical framework underlying the current AI alignment debate has no way to answer this question when AI and the economic model in which it is embedded can so radically change human preferences. We argue that the answer lies in a capabilities-based deliberative ethical framework that views human betterment not as some transcendental endpoint but as an ongoing process of deliberation and exploration which engages and respects multiple points of view. This framework suggests that AI development and use ought to be guided by the principles of enhancing a multi-dimensional view of human well-being, which includes human agency and equitable access to both the process of benefits of AI; co-production of AI among technologists and other members of society; and the democratic shaping of the market forces that in turn shape technology development. These latter two principles may require transparency for AI with privacy for humans as well as a focus on AI as decision support for humans in addition to AI as autonomous agents. These principles of advancing well-being, co-production, and democratic control have been difficult to follow in the world before AI. The challenge now before us is to harness AI to help empower individuals, new policies, and new institutions that can help move closer to these lofty goals.

FIGURES

Figure 1: Societal Influences on Personal Values



Note: Boxes in color and with round edges represent traditional influences. The grey rectangle represents the additional new influence of algorithm-mediated social networks

Table 1: Current social media algorithms and Their Contribution to Human Capabilities

Capability	Definition	Context for Contribution
Life	Able to live to the end of a normal length human life, and to not have one's life reduced to not worth living	*AI might contribute but only in contexts of self-reflection and with additional supplements. Profit-motive of AI deployments not necessarily geared towards improving human life.*
Bodily Health	Able to have a good life which includes (but is not limited to) reproductive health, nourishment and shelter	*AI might provide useful info, but also might interfere; bodily health primarily secured outside AI deployments*
Bodily Integrity	Able to change locations freely, in addition to, having sovereignty over one's body which includes being secure against assault	*AI might provide useful info, but also might interfere; bodily integrity primarily secured outside AI deployments*
Senses, Imagination and Thought	Able to use one's senses to imagine, think and reason in a 'truly human way'–informed by an adequate education. The ability to seek the meaning of life.	*AI can assist, but only with self-reflection and in conditions of effective governance; not a replacement for other pursuits*
Emotions	Able to have attachments to things outside of ourselves	*AI as currently implemented does not tend to foster emotional presence or sustainable attachments*
Practical Reason	Able to form a conception of the good and critically reflect on it	*AI as currently implemented not oriented to fostering critical reflection*
Affiliation	Able to live with and show concern for others; Being treated with dignity and equal worth	*Some affiliations fostered by AI, but risk of polarization and echo-chambers*

Other Species	Able to have concern for and live with other animals, plants and the environment at large	*AI tends to increase engagement with screens and thereby limit our involvement with other species*
Play	Able to laugh, play and enjoy recreational activities.	*AI provides opportunities for play, but also tendency to overwhelm other playful pursuits*
Control over One's Environment	Political – Able to effectively participate in the political life which includes having the right to free speech and association.	*AI can provide fora for expression, but also not translated to long term change*
	Material – Able to own property and the ability to seek meaningful work	*AI as currently implemented helps purchase various property and conduct some types of work*

References

Allen, Danielle, "A New Theory of Justice. Difference without Domination," in Allen, Danielle and Rohini Somanathan, eds., *Difference without Domination: Pursuing Justice in Diverse Democracies*, University of Chicago Press, 2020, pp. 27-58.

Anderson, Elizabeth, *Private Government: How Employers Rule Our Lives (and Why We Don't Talk about It)*, Princeton University Press, 2017.

Ansell, Christopher K., *Pragmatist Democracy: Evolutionary Learning as Public Philosophy*, Oxford University Press, 2011.

Arrow, Kenneth. J., *Social Choice and Individual Values*, Yale University Press, 1953.

Blomkamp, Emma, "The Promise of Co-Design for Public Policy," in M. Howlett and I. Mukherjee, ed., *Routledge Handbook of Policy Design*, Routledge, pp. 59-73, 2018.

Bryant, Benjamin P. and Robert J. Lempert, "Thinking inside the box: A participatory, computer-assisted approach to scenario discovery," *Technological Forecasting and Social Change,* Vol. 77, No. 1, 2010, pp. 34-49.

Carlsen, Henrik, Robert Lempert, Per Wikman-Svahn, and Vanessa Schweizer, "Choosing small sets of policy-relevant scenarios by combining vulnerability and diversity approaches," *Environmental Modelling & Software,* Vol. 84, 2016, pp. 155-164.

Christian, Brian, *The Alignment Problem: How Can Machines Learn Human Values?*, Atlantic Books, 2020.

Defever, C., M. Pandelaere and K. Roe (2011). "Inducing Value- Congruent Behavior Through Advertising and the Moderating Role of Attitudes Toward Advertising." Journal of Advertising, 40:2: 25-38.

Del Missier, Fabio, Mimi Visentini and Timo Mäntylä, "Option generation in decision making: Ideation beyond memory retrieval," *Frontiers in Psychology*, Vol. 5, 2015, pp. 1-16.

Gong, Min, Robert Lempert, Andrew Parker, Lauren A. Mayer, Jordan Fischbach, Matthew Sisco, Zhimin Mao, David H. Krantz, and Howard Kunreuther, "Testing the Scenario Hypothesis: An Experimental Comparison of Scenarios and Forecasts for Decision Support in a Complex Decision Environment," *Environmental Modeling and Software,* Vol. 91, 2017, pp. 135-155.

Helgeson, Casey, "Structuring Decisions Under Deep Uncertainty," *Topoi*, Vol. 39, No. 2, 2020, pp. 1-13.

Guess, A. M., N. Malhotra, J. Pan, P. Barberá, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Dimmery, D. Freelon, M. Gentzkow and S. González-Bailón (2023). "How do social media feed algorithms affect attitudes and behavior in an election campaign?" Science, 381(6656): 398-404.

Jasanoff, Sheila, ed., *States of Knowledge: The Co-Production of Science and the Social Order,* Taylor & Francis, 2004.

Kasprzyk, Joseph R., Shanthi Nataraj, Patrick M. Reed, and Robert J. Lempert, "Many Objective Robust Decision Making for Complex Environmental Systems Undergoing Change," *Environmental Modeling and Software,* Vol. 42, 2013, pp. 55-71.

Köhler, Jonathan, Frank W. Geels, Florian Kern, Jochen Markard, Elsie Onsongo, Anna Wieczorek, Floortje Alkemade, Flor Avelino, Anna Bergek, Frank Boons, Lea Fünfschilling, David Hess, Georg Holtz, Sampsa Hyysalo, Kirsten Jenkins, Paula Kivimaa, Mari Martiskainen, Andrew McMeekin, Marie Susan Mühlemeier, Bjorn Nykvist, Bonno Pel, Rob Raven, Harald Rohracher, Björn Sandén, Johan Schot, Benjamin Sovacool, Bruno Turnheim, Dan Welch, and Peter Wells, "An agenda for sustainability transitions research: State of the art and future directions," *Environmental Innovation and Societal Transitions,* Vol. 31, 2019, pp. 1-32.

Kop, M. (2021). "EU Artificial Intelligence Act: The European Approach to AI." Transatlantic Antitrust and IPR Developments.

Lempert, Robert and Sara Turner, "On Model Pluralism and the Utility of Quantitative Decision Support," *Risk Analysis,* Vol. 41, No. 6, 2021, pp. 874-877.

Lempert, Robert J., Lisa Busch, Ryan Brown, Annette Patton, Sara Turner, Jacyn Schmidt, and Tammy Young, "Community-Level, Participatory Co-Design for Landslide Warning with Implications for Climate Services," *Sustainability,* Vol. 15, No. 5, 2023, p. 4294.

Lempert, Robert J., Steven W. Popper, and Steven C. Bankes, *Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis*, Santa Monica, CA: RAND Corporation, 2003.

Lempert, Robert J., and Sara Turner, "Engaging Multiple Worldviews With Quantitative Decision Support: A Robust Decision-Making Demonstration Using the Lake Model," *Risk Analysis,* Vol. 41, No. 6, 2021, pp. 845-865.

Marchau, Vincent A. W. J., Warren E. Walker, Pieter J. T. M. Bloemen, and Steven W. Popper, *Decision Making under Deep Uncertainty: From Theory to Practice*, Springer, 2019.

Mazzucato, Mariana, *The Value of Everything: Making and Taking in the Global Economy*, Hachette UK, 2018.

McCoy, Martha L., and Patrick L. Scully, "Deliberative Dialogue to Expand Civic Engagement: What Kind of Talk Does Democracy Need?," *National Civic Review,* Vol. 91, No. 2, 2002, pp. 117-135.

National Research Council (NRC) (2009). Informing Decisions in a Changing Climate. Washington, DC, Panel on Strategies and Methods for Climate-Related Decision Support, Committee on the Human Dimensions of Climate Change, Division of Behavioral and Social Sciences and Education.

Office of Science and Technology Policy (OSTP) (2022). Blueprint for an AI Bill of Rights. Washington, DC.

Parson, Edward (Ted), Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant and Nicholas Novelli, *Could AI Drive Transformative Social Progress? What Would This Require?,* UCLA School of Law, Public Research Paper No. 19-49, 2019.

Parker, Andrew M., and Baruch Fischhoff, "Decision-making competence: External validation through an individual-differences approach," *Journal of Behavioral Decision Making,* Vol. 18, No. 1, 2005, pp. 1-27.

Polanyi, Karl, *The Great Transformation: The Political and Economic Origins of Our Time*, Beacon, 1957.

Rahmattalabi, Aida, Shahin Jabbari, Himabindu Lakkaraju, Phebe Vayanos, Max Izenberg, Ryan Brown, Eric Rice, and Milind Tambe, "Fair Influence Maximization: a Welfare Optimization Approach," *Proceedings of the AAAI Conference on Artificial Intelligence,* Vol. 35, No. 13, 05/18, 2021, pp. 11630-11638. As of 2023/06/16: https://ojs.aaai.org/index.php/AAAI/article/view/17383

Rahwan, Iyad, "Society-in-the-loop: programming the algorithmic social contract," *Ethics and Information Technology,* Vol. 20, No. 1, 2018, pp. 5-14.

Rawls, John, *A Theory of Justice*, Harvard University Press, 1971.

Russell, Stuart J., *Human Compatible: Artificial Intelligence and the Problem of Control*, Penguin, 2019

Schon, Donald, and Martin Rein, *Frame Reflection: Toward the Resolution of Intractable Policy Controversies*, Basic Books, 1994.

Schwartz, S. H. (2006) "Basic human values: An overview." 207-208.

Sellers, Charles, *The Market Revolution: Jacksonian America, 1815-1846*, Oxford University Press, 1994.

Sen, Amartya, *Development as Freedom*, Oxford University Press, 2001.

Sen, Amartya, *The Idea of Justice,* Belknap Press, 2009.

Warren, C., A. P. McGraw and L. Van Boven (2011). "Values and preferences: defining preference construction." Wiley Interdisciplinary Reviews: Cognitive Science 2(2): 193-205.