University of Colorado Law School
Colorado Law Scholarly Commons

Publications | Colorado Law Faculty Scholarship

2023

Naïve Realism, Cognitive Bias, and the Benefits and Risks of AI

Harry Surden
University of Colorado Law School

Follow this and additional works at: https://scholar.law.colorado.edu/faculty-articles

 Part of the Internet Law Commons, Law and Psychology Commons, and the Science and Technology Law Commons

Citation Information

Harry Surden, *Naïve Realism, Cognitive Bias, and the Benefits and Risks of AI*, Yale J. on Regul. Notice & Comment (Mar. 16, 2023), https://www.yalejreg.com/nc/naive-realism-cognitive-bias-and-the-benefits-and-risks-of-ai-by-harry-surden

Copyright Statement

Copyright protected. Use of materials from this collection beyond the exceptions provided for in the Fair Use and Educational Use clauses of the U.S. Copyright Law may violate federal law. Permission to publish or reproduce is required.

This Book Review is brought to you for free and open access by the Colorado Law Faculty Scholarship at Colorado Law Scholarly Commons. It has been accepted for inclusion in Publications by an authorized administrator of Colorado Law Scholarly Commons. For more information, please contact rebecca.ciota@colorado.edu.

Naïve Realism, Cognitive Bias, and the Benefits and Risks of AI, by Harry Surden

yalejreg.com/nc/naive-realism-cognitive-bias-and-the-benefits-and-risks-of-ai-by-harry-surden/

**This is the tenth post in a symposium on Orly Lobel's The Equality Machine: Harnessing Digital Technology for a Brighter, More Inclusive Future, selected by The Economist as a best book of 2022. All posts from this symposium can be found here. Further reviews can be found at Science, The Economist, and Kirkus.*

In the *The Equality Machine*, Orly Lobel expertly presents a nuanced view of both the positive and negative aspects of artificial intelligence (AI) technology.[1] This is no small feat, as it has become increasingly popular for commenters on emerging technologies to fall into one of two polarized camps: techno-solutionism or techno-criticism. At its extreme, techno-solutionism tends to lionize technologies like AI and unrealistically position them as easy solutions for the much more complex, systemic issues in society. Conversely, techno-criticism, at its extreme, tends to overemphasize the negative aspects of technologies like AI, either by unduly focusing on potential future problems that may or may not occur, or by disproportionately highlighting edge-cases where a technology is problematic while overlooking other areas where it may be bringing incremental, or significant, societal improvements.

To be clear, scholars from these more emphatic communities have made many valuable contributions to the public discourse around technologies like AI. My intention is not to discount these insights. Instead, I want to reflect on the difficult feat that Lobel accomplished in "The Equality Machine," by engaging with an emerging technology like AI through a lens that is more evenhanded than is typical. It is no easy task, as Lobel has done, to integrate research from both the critical and the technologist communities in order to present a balanced portrayal of the current and future AI landscape, that is neither reflexively critical nor unrealistically optimistic. In reading Lobel's book, I found myself not just appreciating her substantive analysis, but also considering the various cognitive biases that a scholar like Lobel has to work through when attempting to approach a complex and nuanced topic, such as AI from a more balanced viewpoint, rather than from a position that is unjustifiably positive or negative.

In this context, I was reminded of one of psychology's most useful concepts – 'naive realism.' [2] Naïve realism is a cognitive bias – the psychological tendency for our brains to fool us into believing that our personal experiences and knowledge are an accurate representation of the reality of the broader world. It is an interesting fact to ponder that most of what we "know" comes from just a few sources: what we personally experience in the world, what we read or see in media (e.g., such as news, books, television, movies, social media, and the Internet), what we learn from directly talking to others, and how we

cognitively interpret any of this. Naïve realism is the human tendency to over-extrapolate from these quite limited sources of information and believe that this represents the objective reality of the world more broadly. To the extent that we do not fully recognize (or compensate for the fact) that our limited experiences give us but a small and skewed subset of a vastly bigger, more complex, and more diverse world, with many different sources of information and human and cultural experiences, naïve realism functions as a distorting cognitive bias.[3]

Closely related are the concepts of "epistemic bubbles" and the "availability heuristic." Roughly speaking, an "epistemic bubble" refers to human tendency to over-use information sources that arise from the paths of least resistance.[4] They typically arise from information sources that our social peers signal as valid or useful and which flow to us through our ordinary daily activities. More broadly, "Epistemic Bubbles" are a manifestation of cognitive bias known as "selective exposure." As C. Thi Ngugyen describes them, "Epistemic bubbles can form with no ill intent, through ordinary processes of social selection and community formation. We seek to stay in touch with our friends, who also tend to have similar political views. But when we also use those same social networks as sources of news, then we impose on ourselves a narrowed and self-reinforcing epistemic filter, which leaves out contrary views and illegitimately inflates our epistemic self-confidence."[5] This, amplified by the "availability heuristic," which is the cognitive tendency for humans to overweight information that is readily accessible, tends to produce a highly selective, and non-representative view of many issues in the world, including emerging technologies like AI. [6]

Technology itself, not surprisingly, can both ameliorate or exacerbate these cognitive biases. In terms of technology worsening the problem, there is the well-known "extremity bias" effect of social media, in which unusually controversial or extreme events (or viewpoints) tend to be widely propagated and shared at the expense of more representative but ordinary events or moderate views.[7] This is fueled by, and takes advantage of, another cognitive bias – "negativity bias," which is the human tendency to pay more attention to negative information than to neutral or positive information; such disproportionate attention can contribute to a undue sense of anxiety and fear relative to the underlying reality.[8]

More broadly, all of these factors can exacerbate unrepresentative epistemic bubbles with respect to complex issues like AI. This is particularly true for those that are widely discussed over technologies like social media (and the Internet more broadly), where our information sources tend be skewed through selective exposure, algorithmic filtering, and our interpretations of these information sources further distorted through cognitive effects such as extremity bias, negativity bias, and availability bias. Finally, it is difficult to even recognize the lack of objectivity in our world perception due to "naïve realism" and the subtle manner in which our information sources become skewed in ways that do not represent the complex, multi-faceted, and nuanced world at large.

On the other hand, technology also offers new and powerful ways to ameliorate some of these biases that can passively distort our views. Epistemic bubbles, negativity bias, the availability heuristic, naïve realism, and other cognitive biases have always existed in human thinking, and long predate modern communication technology. Prior to the advent of modern communications, people were also constrained by limited access to information, often relying on their community or trusted institutions for news and knowledge, which similarly created a skewed understanding of the world.

Research has shown, with work and effort, one can at least reduce some of the selection effects through various measures, often using leveraging technology, even if one cannot fully eliminate these distortions.[9] Wider access to information allows more of us, in principle, to learn about cognitive biases at all, and then to use this knowledge to recognize that we all are subject to naïve realism and epistemic bubbles, and that our brains are falsely representing our subjective experiences as an objective view of a world. Another affordance of modern technology that can reduce the effect of distorting cognitive biases is the ability seek out a variety of sources of information that are both reliable and that might counter our instinctive, reflexive world views.[10]

Here, Lobel exemplifies the type of diligence required to ameliorate the cognitive biases that can unduly skew one's perception of a technology like AI in either an overly positive or negative direction. In the "Equality Machine," she has done an excellent job of seeking out and engaging with a wide range of literature that is both critical and concerned about AI and the future, and those that recognize the current and potential benefits and gains of AI, while also exploring the various shades of gray that necessarily accompany a topic as complex as AI.

Importantly, Lobel achieves this without detracting from the important contributions of AI critics. These critical scholars of technology play a valuable role in highlighting the risks and issues associated with emerging technologies like AI, including their capacity to perpetuate social inequalities. I personally have found the contributions of critical communities such as the Critical Legal Studies scholars, the Fairness Accountability and Transparency Community (FAT), Science and Technology Studies communities, to be essential in exposing the structural problems and biases embedded in socio-technological systems, and in highlighting diversity problems among those who create and use technology. These voices serve as a necessary counterbalance to some of the unrealistic techno-optimism that often characterizes the private sector's promotion and selling of its technology. Indeed, I myself have at times played the critic in highlighting the problems that AI and other technologies can play in perpetuating social issues, bringing about unwanted externalities, or causing disruptive social change.[11] There are real problems that emerging technologies create that should not be lightly dismissed, such as software that is designed to be addictive, or technologies that can be used to spread misinformation and destabilize democracy.

At the same time, it is equally important not to be reflexively critical to the point that one fails to acknowledge when technology is bringing societal improvements. In this, Lobel plays a masterful role in highlighting such positive aspects of AI. For example, Lobel describes the many ways in which AI has been used to bring to the surface inequalities, using data analysis, that had previously existed in society, but had been hard to detect prior to the advent of AI.

Here, I place her in good company with scholars like Steven Pinker in “The Better Angels of Our Nature: Why Violence Has Declined,” who aim to emphasize and elevate the numerous societal benefits, brought about by social and technological progress, that are today so common and routine that it is easy to take them for granted and overlook the great suffering that existed prior to their development.[12] It is easy for the human mind to conceive of worst-case-scenario problems that AI technology might wreak on society, but it often takes much more mental effort to conceive of the subtle and incremental societal improvements in terms of scientific research, social equality, healthcare, access to information and knowledge, communication, enhancements to the arts and entertainment, communication, standard of living, safety, new types of jobs, that technologies like AI might also bring about.

Overall, Lobel walks an admirable line that realistically balances critique and optimism around technologies like AI, a mode of analysis that I think that more technology scholars should seek to replicate. On the one hand, it is important to acknowledge the reality of those who are suffering today because of technological change, and we do not want to dismiss or minimize their plight. We should not use past technological progress as an excuse not to improve as a society. On the other hand, it is equally problematic to take for granted important progress brought about by technological or social change, to fail to recognize and acknowledge when worst-case-scenarios about technology have not come to pass, or to appreciate incremental but positive changes. In sum, we should strive to be mindful of the cognitive biases that might cause us to have a unrepresentative view of a complex world, and emphatically take unduly inflexible positions one way or another, and take upon ourselves the difficult work it takes to ameliorate them.

With the recent advent of advanced, flexible Large Language Models (LLM) like ChatGPT and GPT4, there is little doubt in my mind that things are moving quickly in the realm of AI. [13] In the 20 years that I have been studying (and programming) in the realm of AI, I have never seen AI advances moving at the rate they currently are.[14] There are many reasons for this acceleration, including better hardware and software, but one significant part is that developers have learned to effectively use AI systems to train (and create software code) for *other AI Systems*. Although the method of using one AI system to train another has been around for some time, the highly effective use of this technique is one of the underlying approaches that propelled ChatGPT well beyond the capabilities of other, similarly sized, contemporary large language models. [15] There is little doubt that this advanced LLM AI technology will prove disruptive in some way to modern society.

However, humans are adaptable creatures, and we tend to adjust to new aspects of technology. We do not know what the future holds, but it is a mistake to conclude that our cognitively-bias-induced worst fears, or more optimistic hopes, are likely to come true. If experience has shown anything, more probably it will be a bit of both, with the neither the worst-case, nor the best-case, AI scenarios coming to pass. What Lobel has shown us is that it takes work to move beyond our inherent cognitive biases, and with effort, we can get a more balanced picture of the positives and negatives of emerging technologies like artificial intelligence. We are not passive recipients of a deterministic technological future; rather, as members of the public, we can and should, take active steps to steer our own society to bring about the future that we want to see.

Harry Surden is a Professor of Law at the University of Colorado Law School and affiliated faculty at the Stanford CodeX Center for Legal Informatics. His primary research area is artificial intelligence and law and he has written numerous articles on the topic (e.g., "Artificial Intelligence and Law: An Overview"). Professor Surden was a professional software engineer prior to entering law.

---

[1] *See* Orly Lobel, The Equality Machine: Harnessing Digital Technology for a Brigther, More Inclusive Future 41 (2022).

[2] *See, e.g.,* Sam Smith, *Teaching Further Education Students the Effects of Naive Realism, to Support Social Development and Mitigate Classroom Conflict*, 37 Educational & Child Psychology 56 (2020).

[3] Of course, naïve-realism, like most other cognitive heuristics, has a beneficial aspect from an evolutionary perspective, as it allows us to get, for many purposes, a "good enough" understanding of the world for many activities, given a world that would otherwise be too complex to handle.

[4] C. Thi Nguyen, *Echo Chambers and Epistemic Bubbles*, 17 Episteme 141 (2020).

[5] Id.

[6] Amos Tversky & Daniel Kahneman, Judgment Under Uncertainty: Heuristics and Biases, in Judgment Under Uncertainty 3, 11 (Daniel Kahneman, Paul Slovic & Amos Tversky eds., 1982).

[7] Leif Brandes, David Godes & Dina Mayzlin, *Extremity Bias in Online Reviews: The Role of Attrition*, 59 Journal of Marketing Research 675 (2022).

[8] *See., e.g.,* Amrisha Vaish, Tobias Grossmann & Amanda Woodward, *Not All Emotions Are Created Equal: The Negativity Bias in Social-Emotional Development*, 134 Psychol Bull 383 (2008).

[9] *See, e.g.,* Tiffany S. Doherty & Aaron E. Carroll, *Believing in Overcoming Cognitive Biases*, 22 AMA Journal of Ethics 773 (2020).

[10] Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. Psychological Review, 111, 781-799.

[11] *See,* Harry Surden, *Values Embedded in Legal Artificial Intelligence*, (2017), https://papers.ssrn.com/abstract=2932333 (last visited Mar 15, 2023), Harry Surden, *Structural Rights in Privacy*, (2007), https://papers.ssrn.com/abstract=1004675 (last visited Mar 15, 2023).

[12] *See, e.g., Pinker, S. (2012). The better angels of our nature.*

[13] *See, e.g.,* ChatGPT, https://chat.openai.com (last visited Mar 15, 2023), ChatGPT, https://chat.openai.com (last visited Mar 15, 2023).

[14] *See, e.g.,* Harry Surden, *ChatGPT, Large Language AI Models, and Making Law More Accessible*, (2023), https://www.harrysurden.com/wordpress/blog (last visited Mar 15, 2023).

[15] Long Ouyang et al., *Training Language Models to Follow Instructions with Human Feedback*, (2022), http://arxiv.org/abs/2203.02155 (last visited Mar 15, 2023).