
Evaluating Metrics for Impact Quantification

Ryan Jenkins and Lorenzo Nericcio



July, 2023

A report supported by the
Center for Advancing Safety of Machine Intelligence

Ryan Jenkins, PhD

ryjenkin@calpoly.edu

Ethics + Emerging Sciences Group

Philosophy Department

California Polytechnic State University, San Luis Obispo

Lorenzo Nericcio, MA

College of Arts & Letters

San Diego State University

Cover photo: This cover has been designed using assets from
[Freepik.com](https://www.freepik.com).

Suggested citation: Jenkins, Ryan and Lorenzo Nericcio. "Exploring
Methods for Impact Quantification." Report funded
by Center for Advancing Safety of Machine Intelli-
gence (CASMI). July, 2023.

Last major revision: June 24, 2023

Version: 0.9

Contents

Executive Summary	1
Project Team	2
Acknowledgments	2
Introduction	4
Preamble	4
Motivation	5
Case Studies	7
Advancing the State of the Art	8
Our Approach: The Human Impact Scorecard	10
The Leap of Faith	10
An Ecumenical Theory of Human Flourishing	13
Identifying Dimensions of Flourishing in the Domain	17
Choosing Metrics to Quantify Impacts on Relevant Capacities	19
Case Studies	21
Case Study: Healthcare	21
Case Study: Recidivism	22
Case Study: Homeland Security	23
Conclusion	23
Background and Rationale	25
Adjacent Approaches in ai Ethics	25
Previous Work on Measuring Multidimensional Impacts	29
Philosophical Challenges	39
The Universality of the Capabilities Approach	45

Some Mature Quantitative Impact Metrics	47
Next Steps.....	51
Sociotechnical Analysis.....	51
Other Species of Harm	52
Works Cited.....	54

Executive Summary

This project proposes concrete metrics to assess the human impact of machine learning applications, thus addressing the gap between ethics and quantitative measurement. Current discussions of AI ethics revolve around fairness, accountability, transparency, and explainability (the so-called “FATE” principles), yet lack concrete metrics for practically implementing or measuring the ethical dimensions of AI. Our report proposes such metrics, defends their philosophical foundations, and illustrates how they can be implemented to facilitate analysis and decision making throughout an organization. We improve upon existing risk assessments of AI, moving us closer to the goal of precise, quantitative assessment of AI’s human impacts.

We outline a universal theory of human flourishing, based on Martha Nussbaum and Amartya Sen’s “capabilities approach.” This theory encompasses broad categories such as environmental health, bodily health, and freedom of affiliation. The approach’s wide scope and “ecumenical” nature allows us to circumvent contentious debates about what constitutes a good life while accommodating a broad array of reasonable views.

Next, we suggest **selecting relevant capabilities that align with the goal of the domain** wherein the AI model is deployed.

Last, we **identify relevant metrics to measure an application’s impact on human flourishing** and propose a “Human Impact Scorecard” that can include both qualitative and quantitative metrics. These scorecards allow for comparisons between applications, thus enabling informed decision-making. We illustrate this approach by applying it to three real-world case studies.

The report, up until this point, stands on its own. If a reader is curious about our methodology, in the latter portion of the report, we explore **the philosophical foundations, adjacent approaches, and previous work** that informs our approach. This analysis reveals several philosophical challenges confronting multidimensional measures of human wellbeing, against which we establish requirements for a successful approach. We show how our approach satisfies these requirements better than any competing approach we know of.

Finally, we discuss **potential extensions of this work**, like accommodating a broader sociotechnical analysis of AI models and addressing other species of harm that might be caused by AI systems.

Project Team

Ryan Jenkins, PhD., is an Associate Professor of Philosophy and a Senior Fellow at the Ethics + Emerging Sciences Group at California Polytechnic State University in San Luis Obispo, and a former Co-Chair of the Robot Ethics Technical Committee of the IEEE. He studies the ethics of emerging technologies, especially AI and data ethics, driverless cars, and robot ethics. His work has appeared in journals such as *Techné, Ethics & Information Technology*, the *Journal of Military Ethics*, and *Ethical Theory and Moral Practice*. He has spoken to broad audiences at VMworld, the New America Foundation, SXSW, and TEDx, and published on technology ethics in public fora including the *Washington Post*, Slate and Forbes. He has served as PI or senior personnel on several National Science Foundation grants to study the ethical, legal, and social implications of emerging technologies, including autonomous vehicles, predictive policing, cyberwarfare, space-based assets, and automated kitchens.

Acknowledgments

I am profoundly grateful to those who have lent their expertise, time, and support to this report. My thanks go to **Lorenzo Nericcio**, whose extensive assistance with authorship and research was indispensable. I am indebted to **Roman Yampolskiy**, whose technical consultation enriched our understanding of AI risk and the plausibility and utility of the case studies featured in this report. **Jezzia Smith**, who conducted research on AI impact assessments and adjacent approaches to measuring AI impacts, provided a solid foundation for our analysis.

The contributions of our dedicated student assistants were invaluable. **Louisa Savageaux**'s research on political decision making, AI for social good, model cards, effective altruism, and her extensive editing played a critical role. **Daniela Flores** deserves mention for her incisive research on legal damages, compensation for the wrongfully convicted, and the emerging research on the intersection of value-sensitive design and AI ethics.

This work was made possible through the generous support of the **Center for Advancing Safety of Machine Intelligence (CASMI)** at Northwestern University. I would like to express my sincere gratitude to **Susanne Gartner** and **Julia Pierce** from the Grants Development and Sponsored Projects Offices

at Cal Poly, whose professionalism and meticulousness shepherded the preparation and administration of the award.

Last but not least, my sincere gratitude goes to **Kristian Hammond** and **Sarah Spurlock** at CASMI. Their guidance, encouragement, and unwavering support throughout this journey have been heartening, as we pursue a holy grail of AI ethics.

Introduction

Preamble

The early Greeks treated mathematics with a mystical quality. Discovering the precise mathematical ratios that underlay musical harmonies—and contemplating the possibilities it opened for understanding the world—intoxicated Pythagoras and his followers.¹ They became transfixed by the promise of quantifying the aspects of experience long thought to be among the most noble—beauty, art, ethics—which had for so long been also the most imprecise, vague, or hopelessly controversial subjects of inquiry. A generation later, in this tradition, Plato calculated that the life of a just person is exactly 729 times happier than the life of an unjust person (Plato 2004, 587e1–4).

Few people accept Plato’s math today. But a kind of “math envy” persists among humanists: the “hard sciences” seem to have a methodological superiority largely in virtue of their ability to present their answers with mathematical precision. And for better or worse, policy-makers and other decision-makers are most comfortable when they can appeal to “hard numbers” to defend their choices. Aristotle lamented as he embarked on his most significant work in ethics, *The Nicomachean Ethics*, that each area of inquiry admits of different degrees of precision (Aristotle 2019, I.3 1094b). The study of the good life has seemed doomed to remain one of the fuzziest domains of inquiry. Its foundations are opaque and controversial; its data are mere *feelings*; any claims about significance or weights must be *subjective*; and so on. As a natural consequence of this, ethicists have often been thought *incapable* of entering into conversations about the humanistic dimensions of technology’s progress beyond offering hand-wringing alarmism or naive Luddism.

The proliferation of technology—and, especially, automated decision-making tools—has made revisiting this assumption more urgent. It is time to redouble our efforts to investigate whether it is possible to develop metrics to measure the impact of technology on human life. Few technologies have

1 Urban legend has it that when one of Pythagoras’s followers discovered irrational numbers, he was drowned at sea, such was the seriousness of this threat to the mythos of a *perfectly rational* cosmos (Huffman 2019, §3.4).

attracted as much attention and generated as much anxiety as artificial intelligence, which is the focus of this project. Few would name a technology expected to be more impactful in the next decade or so. While it's uncontroversial that artificial intelligence will have profound, wide-ranging, transformative effects on societies, economies, and human lives, our grasp of those effects remains stubbornly nebulous. This project seeks to fill a gap in our understanding and planning by developing metrics to quantitatively measure the human impacts of machine learning applications and to thoroughly document the rationale, process, and hurdles in doing so to lay the groundwork for future refinements.

Motivation

Much of the conversation around “ethical artificial intelligence” has come to rest around a stable consensus. Artificial intelligence must be fair, accountable, transparent, and explainable, exhibiting the so-called “FATE” principles. But the need to concretize and operationalize these principles is so far unmet. One expert interviewed in the course of this project remarked speculatively: “machine learning is fundamentally about numeric optimization, so if you have reliable numeric criteria for moral concerns, you could write a function to minimize or maximize those quantities, and that could be incorporated specifically into the paradigm of ML fairly readily.” Thus, there is a need not just for methods for “putting principles into practice,” but to develop, in particular, quantitative metrics which can be used to tune machine learning applications to be sensitive to ethical concerns.

Both Deloitte and the National Institute of Standards and Technology outline guiding principles and frameworks for the identification, management, and avoidance of the possible risks and negative impacts of AI systems. NIST and Deloitte together agree that the state of the art in anticipating and managing AI risk is qualitative rather than quantitative. Two reasons for this may be either that the field is too young and that this work is ongoing, or that it's simply impossible to develop quantitative metrics that work across domains.

Deloitte's report on AI and risk management primarily argues for the implementation of effective risk management in Financial Services (FS) firms that are adopting Artificial Intelligence systems. While the authors repeatedly emphasize the importance of companies identifying the existing risks of the AI systems, they fail to propose any concrete suggestions for how firms should go about recognizing these issues or measuring their impact. However, this piece does offer strong support for the usefulness of a metric that measures the impacts (both positive and negative) of AI systems in the workplace. If a firm wishes to embed AI into their everyday practices, they must understand the

impact that this shift will have on the culture of their company and their talent strategies. In a case such as this, it would be helpful to have some sort of metric that would indicate the human and economic impact of AI implementation (Deloitte 2018).

Similarly, NIST's AI Risk Management Framework (AI RMF) puts forth a set of guidelines that is intended for voluntary use in addressing risks in the design, development, use, and evaluation of AI products, services, and systems. The primary goal of the AI RMF is to identify and manage AI risks and impacts, which underscores the risks of biases or inaccuracies that certain AI systems have demonstrated, and the need for a metric to concretely define the possible damage this could cause along with the probability of these effects materializing. AI RMF uses a taxonomy of three separate classes that identifies characteristics that should be considered when attempting to measure the risks of AI systems (technical characteristics, socio-technical characteristics, and guiding principles), but fails to provide any concrete metrics regarding human impact (Tabassi, n.d.).

AI NOW concurs in their report, "The Social and Economic Implication of Artificial Intelligence Technologies in the Near-Term," which is a summary of a public symposium hosted by the White House and New York University's Information Law Institute in 2016. In the report, AI Now makes several recommendations for developers to use at milestones points in the production, use, governance, and assessment of AI systems to address near-term obstacles and opportunities created by the increased use of AI in social and economic domains. Many of these recommendations highlight, specifically, the need for quantitative impact metrics:

- Recommendation #3 encourages companies to support research to develop a method of measuring the accuracy and fairness of AI systems during the design and deployment stage; and further advocates for research that addresses and measures AI errors and harms of AI *in situ*.
- Recommendation #5 suggests that stakeholders support research regarding methodologies for assessing and evaluating the social and economic impacts of AI systems in real-world contexts. Due to the current lack of a comprehensive method for measuring the social and economic impact of AI systems, these systems are integrated into current social and economic domains with no way to calibrate or measure their impact. Research must be conducted to identify methodologies to fully understand (including measuring) an AI system's impact.
- Finally, recommendation #8 also articulates the need for metrics that measure the *human* impact of AI systems, faulting popular codes of ethics as insufficient. More comprehensive professional codes would include the

responsibilities that AI creators have to those who will disproportionately experience the adverse impacts of AI systems, and this in turn would require some metrics of negative human impact.

Case Studies

These case studies further illustrate the utility of metrics to measure human impact, and their value in facilitating decisions throughout the machine learning development pipeline. Each of the people in these case studies is tasked with choosing between multiple ML applications to implement, or else is identifying the optimal place to intervene to minimize the negative impacts of ML applications. But each of them is stymied by a lack of data, specifically, about the expected outcomes of deploying their systems. What they need to help them decide is a metric that can accurately, concretely measure the impacts of ML, and facilitate apples-to-apples comparisons between different models, or different interventions.

- A **hospital procurement officer** is tasked with procuring a new AI system to help doctors diagnose diabetic retinopathy from fundus (retina) scans. They choose between two technology vendors, A and B, each of whom offers a product that they claim can diagnose diabetic retinopathy with a high degree of success. However, these systems have different rates of false positives and false negatives. In order to decide which software program is better, the officer must decide which is more costly: a higher rate of *false positives*, which would lead to over-treatment and wasted man-hours, or a higher rate of *false negatives*, which would mean that some patients whose diabetic retinopathy could have been caught earlier will instead have later, more costly, and less promising interventions.
- A **director at a technology firm** is testing a new machine learning model which predicts the risk of recidivism for recently convicted criminals. The technology is intended to be used as part of the parole decision-making process, such that parole boards will have access to the verdicts of the ML model alongside other information about the convict's history. The director is interested in tuning the model to optimize the human impacts, to reach the optimal equilibrium in the model of false positives and false negatives, while also optimizing the distribution of engineering resources within their own firm. This requires, for example, deciding how many full-time equivalent (FTE) employees to deploy on model training versus on data ingestion, since both could improve the accuracy of the model.
- A **procurement officer working for the Department of Homeland Security (DHS)** is deciding between two facial recognition systems, Prod-

uct A and Product B, to use for authentication in a secure facility. DHS will err on the side of security, meaning that the costs of false negatives are much higher than the cost of false positives. Product A boasts a much more accurate model for facial recognition, but requires much more detailed imagery of the subject's face and additional biometric information to authenticate the subject. Product B is less accurate overall, raising the specter of false negatives, but it can operate with lower resolution imagery.

Possible applications of such quantifiable metrics include: informing the initial decision of whether to implement AI in a specific environment, comparing AI with the status quo to understand marginal impact and cost-benefit calculus; identifying potential interventions in the development and deployment pipeline; prioritizing different kinds of regulations, targeted at the “worst” species of AI failures, and measuring the effectiveness of those regulations; determining damages or remedies for those affected by AI failures; and, finally, attaching a moral cost to false positives and false negatives.

Advancing the State of the Art

Quantifying the damage that could be caused by a specific machine learning (ML) application presents a daunting challenge. As one expert we spoke to observed, “we cannot say *this* kind of output from *this* AI is liable to cause *this* kind of damage.” This helps explain why qualitative analysis remains the state of the art in AI risk assessment. When dealing with intricate systems whose complete characteristics are yet to be identified, and whose effects on users—also not fully characterized—are highly variable, all we can provide are qualitative risk appraisals such as “low, medium, or high.”

Ideally, we would be working with a well-understood, fully characterized model that has direct effects on an individual, whose preferences and wellbeing are in turn clearly understood. The closer we can get to this instrument, the more accurate and concrete our risk appraisals become.

Consider a case where the values at stake are somewhat universally accepted, such as medicine, whose goals are producing longevity or quality of life for patients. Even here, examining “ML for medical applications” would be far too broad for effective assessment. Instead, one would need to narrow the focus significantly—from medicine broadly, perhaps to cancer, further to measure impacts on skin cancer, and even further to a specific *type* of skin cancer with a well-understood prognosis and effects on life expectancy.

For situations where the value of outcomes isn't as clear-cut as in medicine, the task is even more challenging. Take the example of machine learning used

to flag or censor abusive or hate speech in Facebook comments (Facebook 2020; Nieva 2018). On one hand, proactive censorship could prevent suicides resulting from online abuse (Perrigo 2019; Simonite 2019). On the other hand, it might infringe upon freedom of speech due to false positives, and if the model performs differently in different languages, it could potentially undermine political processes in certain cultures, offending their sense of democracy or participatory rights.

The difficulties lie in two main areas. Firstly, the impact of models on different people or societies is incredibly variable and cannot be fully anticipated or retrospectively explained. Secondly, the types of impacts these models may have, e.g. on suicides but also the dynamism of a democratic society, seem impossible to compare directly—philosophers say they are *incommensurable* (Hsieh and Andersson 2021).

However, the law of large numbers offers a potential solution—by using averages or expected utilities, we can draw more confident generalizations across populations. While the state-of-the-art risk assessment rests on broad generalizations about the potential harm—demonstrated by the prevalence of “low, medium, high” risk evaluations—we propose refining this approach. By incorporating a more granular analysis of the magnitude of risk, and introducing a system that considers a variety of potential impacts, we can significantly enhance the current state of the art. These improvements bring us closer to the holy grail of concrete and quantitative assessment of the human impacts of machine learning.

Our Approach: The Human Impact Scorecard

The Leap of Faith

Embarking on an ethical framework inevitably requires a leap of faith. This inherent complexity in ethical decision-making often leads us to rely on our intuition, as suggested by WD Ross, and that even with perfect clarity on the salient considerations and their relative weight, all we can do is “to study the situation as fully as I can until I form the considered opinion (it is never more)” (Ross 2002, 19). While philosophy can help clarify the relevant dimensions of the problem before us, and examine arguments for their inclusion or exclusion, there is an ineliminable need for judgment and intuitive weighting in the final calculus. We shouldn’t expect more precision than this.

This helps explain why the project we’re undertaking here is, as far as we can tell, unprecedented. We seek to bridge this gap and to not just operationalize but to quantify ethical considerations, dramatically reducing the role of intuition in ethical decision making and replacing it with a quantitative element that will facilitate extraordinary confidence in ethical analyses. This has been seen as a holy grail of some philosophers. Still, it has scandalized others who think it’s not just impossible but a downright offensive suggestion.²

Finally, the analysis is not the decision. Even with the most sophisticated instrument to offer clarity, the final decision—whether we see it as a leap of faith or a matter of judgment—is always there. There is always the possibility to reassess and modify the tool itself as part of a process of reflective equilibrium (Daniels 2003; Raz 2008; Cath 2016). And still, such a tool is useful not just for identifying the single best option but also for narrowing down the set

2 There are some who would not feel this sense of scandal at the thought that ethics can be quantified. One obvious example is a person who adopts a monist view of value, like a hedonist utilitarian. However, this viewpoint is not obviously any easier to operationalize—it would require quantifying the value of every consequence of our actions in terms of happiness—and it isn’t obviously more plausible. Indeed, many people find the very simplicity of utilitarianism suspicious.

of acceptable options and highlighting relevant considerations. This enables a more thorough examination of the remaining options.

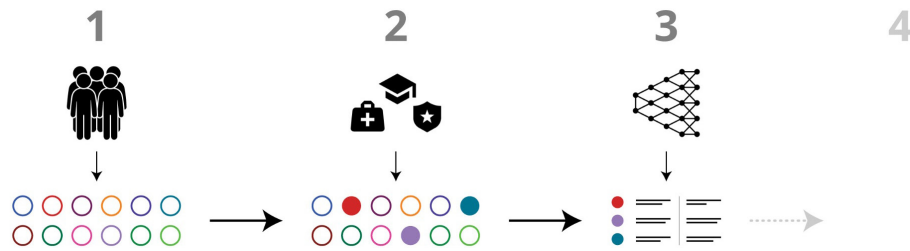


Figure 1. The “master” process we propose for quantitatively evaluating the human impacts of machine learning models. Step 1: Generate an ecumenical theory of human wellbeing to begin from a foundation that can accommodate and acknowledge the broadest range of human impacts. Step 2: Select the dimensions of human wellbeing that will be relevant to the evaluation, depending on the domain in which the model is deployed. Step 3: Using the model as input, evaluate and concretize the human impacts of the model according to mature metrics that quantify impacts in the domain-relevant dimensions. Step 4: Use these quantitative evaluations as inputs into a multi-criteria decision analysis (MCDA) method. (This last step is ultimately outside of the scope of this project and is not illustrated here.)

Our approach to model assessment places a responsibility on those evaluating new models. It requires them to present a strong case concerning specific capabilities and kinds of harms and benefits. Even a qualitative rationale, or a basic checklist that prompts people to think in certain ways, can be valuable. We know that asking individuals to consider their moral reasons often makes them reflect—and may change their behavior (see, among many others, Evans 2008; Craigie 2011; Greene et al. 2001; 2004; Greene 2009).

Our attempt to quantify ethical considerations represents a bold, if challenging, initiative. The process may be fraught with obstacles and disagreements, but it also has the potential to enrich our understanding profoundly. We welcome constructive criticism, and hope this initiative will stimulate discussions—and, in the end, bring us closer to an operational and quantifiable measure of ethics.

However, by carefully approaching the problem of impact quantification, we have developed a novel instrument which offers the precision of quantified decision-making informed by a *pluralistic humanism*. Here, we spell out this view.

Good human lives are a composite of *capabilities*, each of which contributes to an overall state of flourishing.³ Such capabilities include access to healthcare,

3 One technical point to note is the nature of *capabilities* (Robeyns and Byskov 2023, §2.1). Capabilities are, as Sen puts it, real freedoms. They are things that one has the means and ability to do—even if they aren’t presently doing it. Someone may enjoy a freedom even if they choose not to exercise it. For example, someone might choose to never criticize their

ability to play and find daily enjoyment, freedom of religious expression, and others detailed more thoroughly below. Each of these capabilities is in principle measurable or assessable using different quantitative or qualitative measures.⁴ And, consequently, actions, policies, or innovations can in principle be assessed based on their impact on any of them. The particular capabilities that are relevant to each application of machine learning will differ by domain. For some domains, such as medicine, the capabilities of *health* and *physical safety* will be paramount, along with distributive justice. For other domains, such as education, the relevant capabilities will include functional literacy, economic opportunity, and practical rationality, i.e., the ability to reason about the good life. In the abstract, our approach is composed of three claims:

1. There is a set of capabilities that serve as *objective measures* of human wellbeing. That is, for any human being's situation, there are a set of capabilities one may identify whose fulfillment will guarantee a happy, *good* life, and whose lack will cause suffering or languishing.
2. There are components to each of these capabilities which can be understood to indicate how well satisfied each may be in any case. For example, good air quality and good water quality may be compo-

government, and yet still score highly on the freedom of speech capability metric if they live in a state that allows them to do so without fear of persecution.

Capabilities can also be understood as *doings* or *beings*. That is, in addition to the *actions* that someone is capable of doing, there are also *states* that they must be in for them to adequately satisfy a capability. These finer distinctions are not critical to the metrics we evaluate above except insofar as some of the metrics we discuss measure *freedoms to perform actions* that a person may have (like access to outdoor recreation), or they might measure a *being* (like subjective assessment of daily mood). In either case, the metric is useful: it can still provide evidence for the presence of the capability for which it is an indicator.

Some metrics may be more complicated. For example, consider loneliness. Suppose that the prevalence of a certain kind of algorithm has introduced a higher degree of loneliness among a population. This is plausibly the case for young adults and teenagers and social media use; see (Twenge et al. 2021). Young people who feel lonely as a result of social media prevalence may still have the freedom to engage in social activities, at least in principle. But their *frequency* in engaging in them is limited—this is the very thing discovered by such studies. In such a case, it is plausible to assert that, if a trend over time indicates that an AI's impact is positively correlated with the reduction of a kind of *doing* that serves as an indicator for a kind of capability, then that AI's impact has reduced that capability.

4 See, for example, concurrence from Liu et al., “There cannot be one general rule as to which fairness criteria provides better outcomes in all settings” (2019, §3.2) and, “This is consistent with much scholarship that points to the context-sensitive nature of fairness in machine learning” (2022, §6).

nents that accurately pick out cases where the capabilities in the domain *environmental health* are well satisfied.

3. The components in each capability, in each domain, can be assessed: the density of pollutants in the air may be an accurate and consistent way to measure air quality, and so offer some insight into the environmental health capability, demonstrating the degree to which that capability is satisfied. This feature, the quantification or assessability of relevant capabilities, is the operationalizable feature of the theory.

At the operational level, then, our aim is to develop a scorecard based on a catalog of established metrics for each capability. We call this the **Human Impacts Scorecard**. This involves three steps for a given deployment of machine learning:

1. Identify the relevant domain into which the application will be deployed. Identify the goals and values of that domain.
2. Identify capabilities which most closely track the goals and values of that domain.
3. Identify metrics to measure according to the capacities in that domain.

In some cases, this may yield a quantitative output facilitating a direct comparison between alternatives. Often, however, the process will yield, at best, a plurality of qualitative judgments. And so it is at this stage that the qualitative component of our system enters the decision-making process. Before that can be properly understood, we must first outline the precise way that steps (1)–(3) in our theoretical framework operate. That is covered in depth in the following sections.

An Ecumenical Theory of Human Flourishing

One of the best-regarded theories of wellbeing is the capabilities approach developed by Martha Nussbaum and Amartya Sen, a collaboration between moral philosophy and economics, with an eye specifically toward developing a constellation of subjective and objective measures of wellbeing that are amenable to concrete measurement.

As noted above, the capabilities approach emphasizes two components: the promotion of people's wellbeing, and their capability to access the basic conditions necessary for wellbeing (Nussbaum 2008). Further, capabilities are categorized into different broad determinants, like *environmental health*, *bodily health*, or *freedom of affiliation*. Understanding there to be a core set of universal

values allows policymakers, engineers, and other decision-makers to clearly assess how beneficial a given course of action might be. The end-goal is a holistic form of wellbeing, and nurturing these capabilities furthers progress toward that end-goal.

Among academic ethicists, the capabilities approach may be controversial: a strict utilitarian may think it adds extra steps to an otherwise clean utility calculation; a staunch human rights advocate may think that certain rights ought not be violated regardless of the good outcomes that may be produced thereby. However, such edge-cases are not a matter of concern for our framework. We remain *theory agnostic* and seek to develop an account of the good that is *agreeable to most* across theoretical lines. (Plausibly, protecting core capabilities will maximize wellbeing and protect essential human rights anyway.) So, without making contentious theoretical claims, we are most interested in identifying dimensions of human impact. **The capabilities approach affords us the necessary depth and breadth to provide such an ecumenical and broadly acceptable theory.**⁵

We take it as an endorsement of this view of human flourishing that the UN's own approach to measuring the success of their development investments is based on the same fundamental conceptualization of human flourishing. In 2015, the UN identified 17 goals for sustainable development, intended to measure the degree to which lives are improving in developing nations. Such goals include: quality education, reduced inequalities, affordable clean energy, protection of terrestrial and aquatic life, and strong institutions. Each of these goals can be understood straightforwardly as furthering the kinds of human capabilities that Nussbaum and Sen contemplate.

Finally, we will outline the capabilities that serve as indicators for human flourishing. Following Nussbaum and Sen, we identify these domains as those most essential for the support of flourishing human lives:

1. **A full life:** the ability to live a human life of natural length without the threat of an early death or pain so severe as to make life no longer worth living.
2. **Bodily health:** access to shelter, food, and medicine sufficient to keep one's body working and healthy.

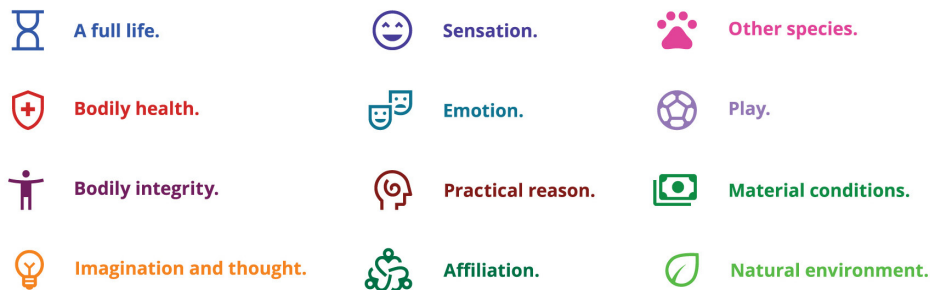
5 As Nussbaum writes of the view: "It is explicitly introduced for political purposes only, and without any grounding in metaphysical ideas of the sort that divide people along lines of culture and religion. As Rawls says: we can view this list as a 'module' that can be endorsed by people who have very different conceptions of the ultimate meaning and purpose of life; they will connect it to their religious or secular comprehensive doctrines in many ways" (Nussbaum 2008, 15).

3. **Bodily integrity:** freedom of movement, freedom from assault, freedom from sexual violence, and reproductive freedom.
4. **Imagination and thought:** the ability to fully use and express one's imaginative capacities, to learn and to reason, to become educated in the areas of one's interest, to produce writing or art aligned with one's interests, and freedom of religious expression.
5. **Sensation:** the ability to pursue pleasurable experiences of different kinds and to avoid non-beneficial pain.
6. **Emotion:** the ability to access the full range of human emotion, from grief to joy, unencumbered by fear and anxiety; also the ability to form relationships with others necessary to experience such emotions.⁶
7. **Practical reason:** the ability to determine the components of a good life for oneself and critically reflect and plan for such a life.
8. **Affiliation:** the ability to live in relation with others freely and by one's choosing, including the political protection of freedom of association; being treated with dignity and respect politically and in social situations, and freedom from discrimination based on gender, sexuality, ethnicity, religion, or national origin.⁷
9. **Other species:** the ability to live with care for animals, plants, and other organisms in the environment.

6 In further discussion below, we merge *bodily health* and *bodily integrity*, as well as *sensation* and *emotion*, as, in practice, the metrics used to measure these goods are very similar.

7 Concerns that AI could be little more than a thinly veiled form of discrimination are popular. In the guise of AI, discrimination takes on the form of *statistical group discrimination* (Gomez, 2018). Statistical group discrimination is defined as negatively impactful, disparate treatment against socially salient groups based on statistically relevant facts (Lieppert-Rasmussen, 2013). By this definition, a socially salient group is one whose membership is important to the dynamic and structure of social interactions across a variety of social contexts (Gomez, 2018). In particular, groups that are defined by properties that are protected by anti-discrimination laws are socially salient. There are three dimensions to statistical group discrimination: generic discrimination, group discrimination, and statistical discrimination. In general, generic discrimination is disadvantageous differential treatment that occurs when X treats Y poorly in comparison to others due to Y possessing property or perceived property P. Group discrimination requires X to generically discriminate against Y and that P is the property of belonging to a socially salient group, which ultimately makes people with P worse off relative to others. X statistically discriminates against Y if X group-discriminates against Y and P is statistically relevant or X believes P is statistically relevant (Lieppert-Rasmussen, 2013). For example, an employer engages in statistical discrimination if they do not hire a highly-qualified woman because the women they have already employed have a higher probability of taking parental leave (Gomez, 2018).

10. **Play:** the ability to enjoy hobbies, humor, recreation, and leisure.
11. **Control over one's material conditions:** applicable both to (a) one's social and political environment—the freedom to participate in political choices that affect one's polity—and (b) material control of one's belongings, employment, and living situation sufficient for comfort and happiness.
12. **Healthy natural environment:** the ability to exist in an environment free from excessive pollution and the ability to enjoy the benefits of healthy ecosystems and natural scenery; knowledge that the environment in which one lives will be healthy and liveable into the future (M. Nussbaum 2000; Sen 2008).⁸



Each of these capabilities is broad. That presents challenges—identifying the best ways to assess satisfaction of capabilities may be difficult—but not without immense benefit: the fact that each domain is broad permits multiple, culturally-sensitive ways to satisfy each capability. And so, as explained previously, this approach is sensitive to the great variation that exists across cultures and domains of human activity, while still offering the grounding power of an objective list of components of a good life.

Identifying Dimensions of Flourishing in the Domain

The next step is determining what the *relevant dimensions* of flourishing are in any given domain. Consider the case studies offered at the beginning of this document:

⁸ This capability is our only addition to Nussbaum and Sen's list; we felt that environmental health was particularly morally important in a way not sufficiently emphasized in the other capabilities. Note that we also changed the exact wording in the other capabilities to indicate features of each most relevant to our aims in this project, but that they otherwise align with the original list.

- A **hospital procurement officer** needs to choose between two AI systems for diagnosing diabetic retinopathy. They must consider the trade-off between false positives and false negatives to determine which is more costly: over-treatment due to false positives or delayed interventions due to false negatives, which have higher costs and poorer outcomes.
- A **technology firm director** is testing a machine learning model to predict recidivism risk for parole decision-making. They aim to optimize the balance between false positives and false negatives while efficiently allocating engineering resources, such as deciding the number of employees for model training and data ingestion, to improve accuracy.
- A **DHS procurement officer** must decide between two facial recognition systems for secure facility authentication. While Product A has better accuracy, it requires more detailed imagery and biometric data, while Product B has lower accuracy but can operate with lower-resolution imagery. Given the twin emphases on privacy and security, the officer must weigh the costs of false positives against the potential drawbacks of lower accuracy.

For each domain, some dimensions of flourishing will be more relevant than others. In order to ensure that this problem is tractable, it's best to choose cases that clearly occupy a single domain. Analysis should be driven by the relevant goals and values of the domain. For example, in the medical domain, we might be most interested in quality of life, life expectancy, and health disparities or health justice. In criminal justice, we might be most concerned about political liberty, job opportunities, etc. In each domain, we can select down the relevant dimensions of flourishing according to which a machine learning application should be evaluated.⁹

Our theory of domains is inspired by the neo-Aristotelian conception of *domains of practice*—which we call *domains* for short—which is popular among contemporary neo-Aristotelians like Walzer (2010) and MacIntyre (1988; 2007). The domains of practice literature also borrows heavily from the work

9 Note that the method we choose to individuate “domains” is consequential. Even across medical sectors, it may be vague what counts as *transparency* or when accuracy ought to outweigh distributive fairness. Moreover, the domain of “medicine” itself can be spliced further and along different dimensions. For example, some medical professionals specialize in the use of certain tools and techniques, such as radiology or endoscopy; some specialize in parts of the body, such as cardiology or neurology; some specialize in age groups, such as pediatrics or geriatrics. Our domain-level approach is motivated by an attempt to split the difference between evaluations of AI that are too generic to be actionable, on the one hand, and those that are too specific to be portable or generalizable, on the other hand. For much more on this point, see Jenkins et al. (2022).

on *practice-dependence*¹⁰, especially within the global justice debate which followed in the wake of Rawls (James 2005).¹¹

Our claim is that domains each furnish a characteristic good to society. This good is, in turn, relevant to some of the capacities of human wellbeing outlined above.¹² The paradigmatic goals of the domains discussed here tend to be *intrinsic goods*, that is, things that are desirable for their own sake. (The domain of medicine provides patients with health; the domain of journalism provides readers with truth or understanding about the world; and so on. These benefits have been seriously entertained as intrinsically valuable.) Accordingly, each of the capacities outlined above, we suggest, is an intrinsic component of a good human life, and which cannot be further reduced to another capability or explained fully in terms of that capability.

We define a goal as “an outcome we hope to accomplish in a domain.” When we use the word, “goal,” we mean it *aspirationally* as opposed to *descriptively*. Identifying the goal of the domain is accomplished in conversation with the practitioners in the domain, understanding what they take themselves to be doing and how it may be different from adjacent domains. We are not trying to describe people’s actual motivations, because they might be motivated by fame, reputation, money, vengeance or any number of less savory things. Instead, it will be more helpful to ask questions such as:

- What are people within this domain hoping to contribute to society?
- How do the people working in this domain praise themselves, e.g. in their advertisements, award ceremonies, or public statements?
- What benefits do consumers, users, or broader society expect these domains to furnish?

The more seasoned practitioners in a domain should be able, upon reflection, to provide some explanation of and justification for what they are doing with their lives. Their motivations will likely resonate with some of the capabilities outlined above more than others. Still, consensus would be a quixotic goal and

10 See especially James (2005), whose paper was an important catalyst for the recent flurry of scholarship on practice-dependence. See also Jubb (2016), Erman and Möller (2015) and Erman and Möller (2016).

11 Much of the discussion of domains that follows is taken from Jenkins et al. (2022), including a good deal of verbatim material.

12 In fact, this criterion also serves to separate domains from one another. How do we distinguish, for example, between literature and journalism? One of these is supposed to deliver the benefit of helpful information about current events, i.e., to function as “the first draft of history.”

we should be satisfied if we can arrive at answers that *most reasonable practitioners* could accept and that enjoy *wide endorsement*. Beitz is especially helpful here: what we seek is “a facially reasonable conception of the practice’s aim [and values] formulated so as to make sense of as many of the central normative elements as possible within the familiar interpretive constraints of consistency, coherence, and simplicity” (2009, 108).¹³ This process often benefits from consulting the documents promulgated by a domain’s professional bodies, which explicitly lay out a profession’s aspirations and values.¹⁴

Choosing Metrics to Quantify Impacts on Relevant Capacities

Finally, we identify relevant metrics for measuring the impact of the application on the selected domains of human flourishing. Which of these impacts can be most reliably measured? For example, health outcomes could be measured in QALYs. Environmental outcomes may be measured in environmental particulates in the air, or tons of CO₂ emissions averted.

-
- ¹³ We are buoyed by the success of this method in some domains, for example, in the history of the professionalization of journalism, and the coalescence of journalists worldwide around a broadly shared understanding of the goals and values of their work. See Deuze (2005), which traces the history of journalism’s self-perception and the formation of its professional identity, which is “kept together by the social cement of an occupational ideology” (2005, 442). See also Weaver (1998, 456), who argues that the late-20th century “consolidation” of journalistic values even stretched across national borders.
- ¹⁴ In the sociology of professions, authors often take the existence of a code of ethics, which articulates shared understandings and expectations of appropriate behavior, to be crucial for professionalization. See Wilensky (1964) for a classic treatment, and Abbott (1991) and Hall (1988) for other classic discussions of the ‘process’ model of professionalization. See Forsyth and Danisiewicz (1985) for a literature review and defense of alternative theories of professionalization. Still, professions are narrower than occupations, and still more narrow than domains as we understand them. But for those domains containing mature professions, this task is easiest.

Some dimensions of impact will be more difficult to quantify, e.g. those dealing with the mental life and mental functioning of users and broader society. In those cases, qualitative measures might be all we can achieve. Consider a Human Impact Scorecard like the one in Figure 3, attached to a hypothetical app to train students on LSAT practice problems:



 Imagination and thought.	++ This app coaches users on logic and critical thinking.
 Emotion.	+ This app may improve emotional regulation.
 Practical reason.	- This app uses nudges to gamify behavior, which may undermine autonomy.

Figure 3. A human impact scorecard for a hypothetical app using machine learning to train students on LSAT practice problems.

Even a qualitative rendering like this is likely to facilitate meaningful apples-to-apples comparisons between two apps. See Figure 4 for an example:




 Imagination and thought.	++	+
 Emotion.	+	+++
 Practical reason.	-	0

Figure 4. A human impact scorecard providing meaningful apples-to-apples comparisons between two apps even based solely on qualitative measures of impact.

More impressive and powerful comparisons would involve quantitative metrics, such as in Figure 5, which shows a hypothetical app that directs users to nearby parks when they are on long driving trips.


 Bodily health.	+1 yr	People who spend an extra hour in nature per week add one year to their lives on average.
 Play.	+	People with access to the outdoors report more unstructured and playful time.
 Natural environment.	3 days	People who live close to nature spend on average 3 more days per year outside.

Figure 5. A human impact scorecard providing quantitative and qualitative metrics for a hypothetical app that directs users to nearby parks when they are on long driving trips.

Case Studies

To illustrate specifically how our process facilitates analysis and decision making, we will walk through the three case studies introduced above as part of our motivation for the project. These examples are meant to be lightly fictionalized but realistic.

Case Study: Healthcare

A hospital procurement officer is tasked with procuring a new AI system to help doctors diagnose diabetic retinopathy from fundus (retina) scans. They choose between two technology vendors, A and B, each of whom offers a product that they claim can diagnose diabetic retinopathy with a high degree of success. However, these systems have different rates of false positives and false negatives. In order to decide which software program is better, the officer must decide which is more costly: a higher rate of false positives, which would lead to over-treatment and wasted man-hours, or a higher rate of false negatives, which would mean that some patients whose diabetic retinopathy could have been caught earlier will instead have later, more costly, and less promising interventions.

First, our procurement officer has to consider the goals of the domain in which this application will be deployed. Medicine is one of the domains whose goals are clearest. Medicine aims at the quality of life and longevity of the patient or, simply, their health. From our list of capabilities above, this goal is best captured by the following two capabilities:

- **A full life:** the ability to live a human life of natural length without the threat of an early death or pain so severe as to make life no longer worth living.
- **Bodily health:** access to shelter, food, and medicine sufficient to keep one's body working and healthy.

Once it's clear which capabilities are implicated in this analysis, our procurement officer should seek out mature quantitative metrics for those. We have

chosen QALYs for their popularity and appropriateness in the medical context. Effects on treatment success are also included, since false negatives will delay treatment and reduce the likeliness of positive outcomes. (False positives incur no additional cost in terms of life expectancy or treatment success.) Finally, we include calculations for additional FTE hours that are incurred by false positives and negatives, i.e. the extra time spent reviewing false positives to overturn the verdicts of the model, and the extra time spent playing catchup in response to false negatives once the correct diagnosis is made.

Capability	Metric	Vendor A		Vendor B	
		False Positive	False Negative	False Positive	False Negative
Full life	Life expectancy	—	2 QALYs	—	2 QALYs
Bodily health	Treatment success	—	-12%	—	-12%
Expense	FTE hours	3	6	3	6

Case Study: Recidivism

A director at a technology firm is testing a new machine learning model which predicts the risk of recidivism for recently convicted criminals. The technology is intended to be used as part of the parole decision-making process, such that parole boards will have access to the verdicts of the ML model alongside other information about the convict's history. The director is interested in tuning the model to optimize the human impacts, to reach the optimal equilibrium in the model of false positives and false negatives, while also optimizing the distribution of engineering resources within their own firm. This requires, for example, deciding how many full-time equivalent (FTE) employees to deploy on model training versus on data ingestion, since both could improve the accuracy of the model.

The director considers the relevant dimensions along which to assess the distribution of resources within their organization, given their domain. For the criminal justice domain, these capabilities could include access to a full life and control over material conditions.

Capability	Metric	Model training improvement per additional FTE	Data ingestion improvement per additional FTE
Full life	Life expectancy	2 QALYs	2.4 QALYs
Control over material conditions	Homelessness	2% reduction	3% reduction
	Employment	3% increase	4% increase

The results of this analysis show that FTEs devoted to data cleaning and ingestion will have superior expected outcomes over FTEs devoted to model training and fine-tuning.

Case Study: Homeland Security

A procurement officer working for the Department of Homeland Security (DHS) is deciding between two facial recognition systems, Product A and Product B, to use for authentication in a secure facility. DHS will err on the side of security, meaning that the costs of false negatives are much higher than the cost of false positives. Product A boasts a much more accurate model for facial recognition, but requires much more detailed imagery of the subject's face and additional biometric information to authenticate the subject. Product B is less accurate overall, raising the specter of false negatives, but it can operate with lower resolution imagery.

Capability	Metric	Product A Impacts	Product B Impacts
Full life	Life expectancy	2.1 QALYs saved	1.7 QALYs saved
Bodily integrity	Subjective report of security and safety	+2.3	+1.4
Affiliation	Privacy violations	7/10 discomfort	3/10 discomfort

The analysis above shows that while product A is more accurate, is able to ensure greater security for a greater number, providing them with a greater

sense of safety and security, the benefits come at a much greater cost to privacy. In fact, the analysis reveals that Product A comes at a much greater cost to privacy for a marginal increase in QALYs saved.

Conclusion

Our emphasis is on the democratization of AI risk analysis, with the ultimate aim of enhancing ethical AI practices across various sectors. Our goals, to reiterate, are to produce an instrument that is both flexible and amenable to a variety of domains, while also being relatively quick and easy to apply. This methodology should be user-friendly, enabling a non-expert to apply it within a few hours. Teaming up with domain experts—in the technical aspects of machine learning or those with the relevant subject matter expertise for the domain in question—would undoubtedly enhance the output, resulting in more precise and accurate assessments. Even though that collaboration would be beneficial, it is not required, in our estimation, to provide an analysis that improves upon the status quo in facilitating confident and principled decision making. In sum, this work brings us closer to fully generalizable and practicable instruments to operationalize the ethical development and evaluation of machine learning.

Background and Rationale

Adjacent Approaches in AI Ethics

Algorithmic Impact Assessments

One approach to evaluating the impacts of AI applications that has become popular in both public and private organizations is to generate an *algorithmic impact assessment* (or AIA). AIAs have been developed as a straightforward way for regulators and private industry partners to assess outcomes that may be produced by the implementation of a new algorithm. Facial recognition, targeted ads, automated machines, medical software and more may all be subject to AIAs (Selbst 2021). Assessments may anticipate impacts in domains like: data accuracy, fairness, transparency, accountability, privacy, and security.

Developing an impact assessment usually includes the developers of a model filling out a questionnaire. These questionnaires contain a variety of qualitative and quantitative questions designed to characterize the possible risks imposed by an AI system, as well as what measures may be taken to mitigate those risks. Questions may ask about: the problem the algorithm is built to solve; its method for solving that problem; the stakeholders who may be affected; what the risk to those stakeholders might be; and what strategies are presently being developed by the application's creators to mitigate those risks. The output of the questionnaire is a score (along with a sheet cataloging responses) that characterizes the risk level attending different possible impacts.

Ideally, companies can use the output of the assessment questionnaire to inform their development and mitigation strategies—then, once the algorithm has actually been implemented, the next step is to continue the assessment process using data from its actual impacts (rather than merely forecasted ones).

AIA relies on good faith participation by firms developing algorithms, which is difficult to ensure. There is also some concern that this process does not guarantee any community or stakeholder input—and firms may be resistant to including such input (Selbst 2021). Another concern is that, while this system may deliver a basic impact score, this impact is domain agnostic, and these questionnaires seem to blend together the practical and the moral concerns involved in developing models. Similarly, there is nothing in the process morally grounding the output as morally relevant; thus the decisions about what

impacts to assess, and how to weigh them, are somewhat arbitrary. As a case in point, because a typical questionnaire is domain-agnostic, its ability to guide action is undermined.

About the Data - A. Data Source

24. Will the Automated Decision System use personal information as input data?
Yes [Points: +4]
25. Have you verified that the use of personal information is limited to only what is directly related to delivering a program or service?
Yes [Points: +0]
26. Is the personal information of individuals being used in a decision-making process that directly affects those individuals?
Yes [Points: +2]
27. Have you verified if the system is using personal information in a way that is consistent with: (a) the current Personal Information Banks (PIBs) and Privacy Impact Assessments (PIAs) of your programs or (b) planned or implemented modifications to the PIBs or PIAs that take new uses and processes into account?
Yes [Points: +0]
28. What is the highest security classification of the input data used by the system? (Select one)
Protected B / Protected C [Points: +3]
29. Who controls the data?
Federal government [Points: +1]
30. Will the system use data from multiple different sources?
Yes [Points: +4]

Figure 6. Example from the Canadian Algorithmic Impact Assessment Survey.

Value-Sensitive Design for AI

The methodology of value-sensitive design (VSD) developed by Batya Friedman is a method for proactively considering and centralizing human values in the technology design process—and specifically in information technologies (Friedman 1996; Friedman et al. 2013; Friedman, Kahn, and Borning, n.d.; Friedman, Hendry, and Borning 2017). VSD includes a threefold approach of conceptual, empirical, and technical investigations, ultimately combined in an integrative and iterative method of design. There is emerging work in developing VSD methods for the design of artificial intelligence in particular (Umbrello and De Bellis 2018; Umbrello and van de Poel 2021).

The empirical investigations involved in VSD include a careful look into how the technology could be implemented, what the consequences might be of its various uses or functions, the potential drawbacks of specific features, and other opportunity costs. In theory, this stage would incorporate quantifiable measures to reliably measure these impacts, for example, statistical data that describes patterns of human behavior and assessments that measure the needs and wants of users (Yampolskiy 2019). The aim of this stage is to make use of precise measurements, where appropriate, to improve or refine the design

(Yampolskiy 2019). However, there is little research on the deployment of quantifiable methodologies under VSD.

AI for Social Good

The proposal to deploy “AI for social good” (AI4SG) has attracted significant attention and investment (Hager et al. 2019). One component of this interest is the desire to maximize the impact of AI4SG funding, ensuring the maximum return on investment for each dollar donated. Establishing these priorities requires, in turn, a method for measuring and comparing the likely impacts of different deployments of artificial intelligence.

One framework for guiding AI4SG investments recommends considering the size of the potential impact, implementation feasibility, and opportunity for area synergies of the investment. The “best bets” for AI investment, according to these metrics, include: breadth and depth of impact, implementation potential, potential downsides, and opportunities for synergies (Brockman, Ben et al. 2021). The top areas that score the highest according to this investment scale are point-of-care diagnostic tools for low-resourced medical systems, communication tools that support marginalized communities and languages, and agricultural yield prediction in smallholder-dominated regions (Brockman, Ben et al. 2021). Even with these recommendations, the framework can only generate ratings of 1–4 rather than, say, estimates of lives or acres of rainforest saved. This does not facilitate the more fine-grained comparisons contemplated in the motivating case studies above.

Longtermism and AGI

One approach to the assessment of AI that has received significant attention in the philosophical scholarship is Longtermism, the broadly utilitarian position that future generations’ wellbeing should be weighted equally with the wellbeing of generations alive today. This view is often supported by the Effective Altruism movement. Someone assessing AI on this approach is likely primarily be concerned with either:

- AI’s potential for exponentially increasing the total amount of happiness (say, by facilitating interplanetary settlement or allowing humans to become immortal by uploading our consciousness to computers); or
- AI’s potential for producing a human extinction event—e.g., rogue AI determining all humans must be eliminated, or some such extreme outcome.

There are features of this body of scholarship that are useful to this project. For example, many AI researchers and philosophers in this domain are concerned with *value alignment*, the technical process involved in ensuring AI pro-

grams operate in alignment with values determined by human operators. The technical features of this work may well be important for the process we develop later. However, Longtermism itself does not offer much guidance. This is because its proponents' research is primarily concerned with the long-term scope, and AI's potential to produce exponentially-growing benefits or harms. As such, little of the research in this area touches on the exact ways we might measure the small-scale ethical outcomes of implementing an AI system without those potential consequences. For example, an AI that predicts a convict's likelihood of recidivating is unlikely to end the world or help humanity colonize Mars, and so is of little interest to the Longtermist.

Model cards

Originally pioneered by Google, model cards were created with the intention of surfacing information about an AI model's function and capabilities and rendering it legible. Since then, other large companies that employ or develop AI such as Salesforce and Facebook have taken an interest in this form of documentation. Model cards are documents that accompany AI systems that communicate the ideal forms of output, key limitations, and basic performance metrics of an AI system. Ideally, according to Google, model cards will visualize and express quantitative data related to a model's performance in a comprehensive yet accessible manner while avoiding oversimplifying complex categories and concepts such as race and gender. Google's sample model card for an AI facial recognition software primarily included performance metrics such as precision-recall values (PR) and area under PR curve, disparity in recall, and the effect of facial size, orientation, and degree of occlusion on system performance (Google n.d.). Thus, the cards surface quantitative metrics alongside flagging known issues that may have a moral dimension, such as disparities in performance across racial groups.

Mitchell et al. asserts that model cards are the best way to clarify the intended use of a machine learning system as well as minimize the likelihood of their implementation in inappropriate contexts. Model cards are documents that accompany machine learning models that contain benchmarked evaluation conditions (e.g. cultural, demographic, or phenotypic groups) and intersectional groups that are relevant to the domain in which the system is being used. Characteristics of the model cards include model details, intended use, factors, metrics, evaluation data, training data, quantitative analysis, and ethical concerns. The model details section should provide information about the person or organization developing the model, model date, model version, model type, paper or other resources, citation details, license, and feedback on the model. Intended use covers the primary intended use of the system, the primary intended users, and the out-of-scope uses (e.g. related and similar technology or other contexts of use). In the factors section, model cards should

report system performance across a wide range of relevant and evaluation factors, such as the groups that are present in the data instances, the instrumentation that was used to capture the input of the model, and the environment in which the system is deployed. The metrics, evaluation data, and training data should all be determined based on the model's structure and intended use as well as provide visibility into the source and composition of the datasets used. The quantitative analysis of the metrics, evaluation data, and training data should be disaggregated by the chosen factors and based on the model's performance with respect to each factor (unitary results) and the model's performance with respect to the intersection of evaluated factors (intersectional results).

The last and most subjective aspect of the AI model cards is the ethical concerns. The framework Mitchell et al. set forth for ethical contemplation and decision-making includes consideration of sensitive data, whether the model is intended to inform decisions that are central to human life and wellbeing, the risk mitigation strategies employed during the development process, the possible risk and harms of model usage, and particularly concerning use cases. While a good start, this is far from an exhaustive list of potential ethical concerns that range across domains, and lacks completely—because it does not aspire to offer—quantitative metrics for ethical impacts.

Previous Work on Measuring Multidimensional Impacts

Many other disciplines and projects are faced with similar challenges: namely, wrangling a great variety of potential ethical impacts, concretizing them in a way that facilitates decision-making, and measuring the impacts of those decisions. As part of our initial landscaping research, we analyzed nearly a dozen adjacent approaches to quantitatively measuring multidimensional human impacts in domains such as policy, psychology, and law. We discuss some of the more promising approaches here and discuss other less promising approaches in Appendix 3: Previous Approaches. By reflecting on these approaches, we are able to derive criteria for a more successful method.

Evaluation of Some Methods for Measuring Multidimensional Impacts

From our analysis, we have derived multiple requirements for successful approaches to measuring impacts. First, the method should be **maximally generic and flexible**. Many approaches are tailor-made to specific applications within a domain. For example, imagine a hospital that proposes a policy intervention to minimize falls among patients by conducting risk assessments, providing appropriate footwear and mobility assistance, and so on. Developing an

instrument to measure the success of this intervention would be taxing, but could yield exquisitely precise and useful data for evaluating the intervention within the specific domain, application, and deployment context. But an instrument this specific would be difficult to generalize to other applications, perhaps even among other hospitals. We aim to suggest a method that can be flexibly and easily applied across domains while also facilitating quantitative analysis.

Second, the method should be **practicable and mature**, which is to say it should be amenable to quantification and free from serious philosophical objections. Many approaches are quantitative and even unidimensional. This can be a source of great clarity but, by the same token, these approaches invite controversy and criticism by purporting to boil down the complexities of human experience to a single number. See, for example, the QALY metric (explored more below).

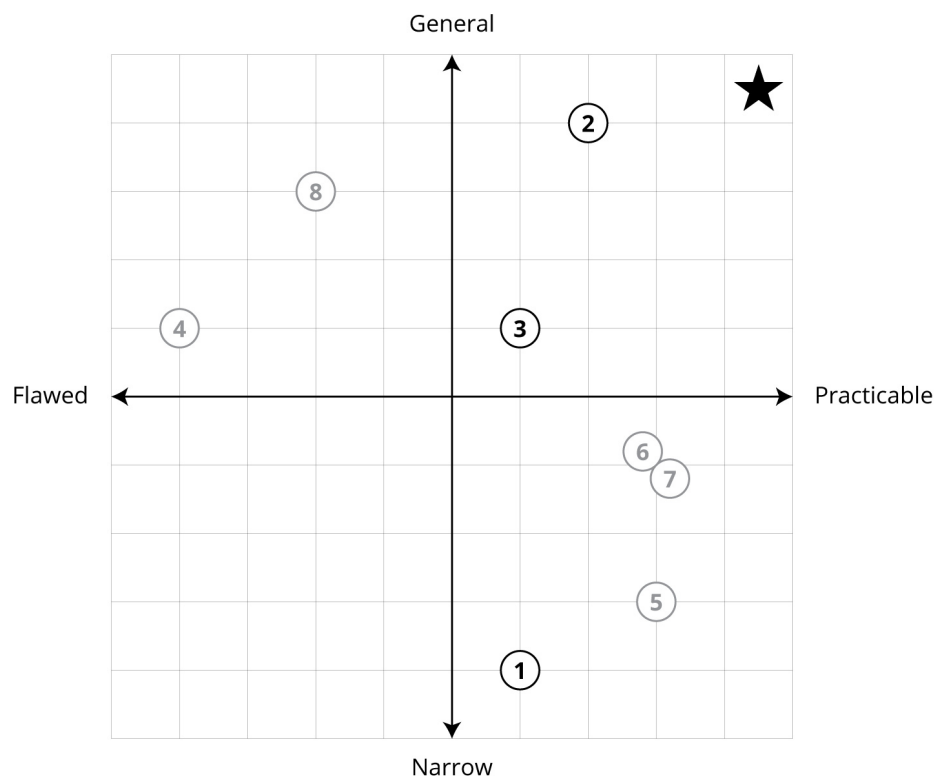


Figure 7. Key to approaches to measuring multidimensional human impacts plotted above.

- | | |
|--|--|
| ★ <i>The Holy Grail</i> | 5. <i>Quality-adjusted life years (QALYs)</i> |
| 1. <i>Subjective wellbeing reports</i> | 6. <i>Effective altruism movement</i> |
| 2. <i>Development economics</i> | 7. <i>Compensating the wrongfully imprisoned</i> |
| 3. <i>Legal damages</i> | 8. <i>Quantitative human rights</i> |
| 4. <i>Political investments and priority-setting</i> | |

Subjective wellbeing reports (1)

One source of data which is a popular tool for psychologists is the *subjective wellbeing report* (SWR), which is gathered by asking a person, usually on a scale of 1–10, how happy they are; how satisfied they are with their life; how safe and secure they feel they are; and so on.¹⁵ These data are promising because they are quantifiable and facilitate direct comparisons, in this case, between countries, which can be used to judge, e.g. their development progress, their policies, or their political structures. However, purely subjective tools have several problems. First, people often have irrational preferences which would not stand up to rational scrutiny. Yet if people are simply reporting their preference satisfaction, then this means that SWRs are tracking the wrong thing. People are also subject to preference cycling and other ways in which their preferences are incoherent or mutually contradictory.¹⁶ Lastly, and most troublingly for our purposes, people are subject to “adaptive preferences,” wherein they can become acclimated to their surroundings, regardless of whether their situation is objectively good or bad. It may be surprising, for example, to find that citizens living under despotic regimes or war-torn countries such as Israel, Kosovo, and the United Arab Emirates tend to rate their subjective wellbeing as very high (Helliwell et al. 2021). Yet it seems that those reports should *not* be taken as reliable evidence that those societies are flourishing.

In responding to this collection of worries, welfare theorists tend to either subject these subjective components to rational constraints (i.e. trying to rule out irrational sets of preferences), or else conjoin SWRs with an objective requirement, such as a measure of political freedom in a country. This is to say, briefly, subjective theories of wellbeing seem to face very serious problems, and the solution to those problems, in one guise or another, tends to be to push the theories toward incorporating objective elements.

Development economics (2)

Development economics concerns itself with approaches to improve economic, social, and political outcomes in developing countries. Thus, the field needs a

15 There is an extensive literature on the philosophy and measurement of subjective wellbeing, owing substantially to the work of Ed Diener. See, for example: (Diener 1994; Diener and Ryan 2009; Helliwell and Barrington-Leigh 2010; Pavot 2018; Lucas 2018; Kahneman and Krueger 2006).

16 A person has “cycling preferences” if, for example, they report the following preferences: A is preferable to B; B is preferable to C; and C is preferable to A. These rankings are mutually inconsistent if we believe that preferences ought to be transitive, in which case this person ought to prefer A to C. Because, instead, they view C as preferable to A, this person seems to be irrational.

method for understanding when exactly it can be said that a country has successfully improved the situation of its citizens. Traditional methods for evaluating development (and so evaluating policy impact more broadly) rely largely or solely on economic evaluations, such as GDP per capita, as proxies for the quality of life of citizens. Amartya Sen and Martha Nussbaum have developed a competing approach to these traditional methods. They suggest that what it is to live well is captured not by an economic metric, but instead by a cluster of capabilities whose enjoyment leads to a good life. Nussbaum and Sen contend that *a country is successfully developing when the wellbeing-associated capabilities of its population are improving*.

Nussbaum and Sen focus on two core views: (1) there exists a set of *capabilities* that are necessary for the pursuit of a good and meaningful life, and (2) that the best way to understand freedom, and so to morally evaluate a country's state of development, is by measuring people's capabilities. On this view, a capability is something that a person must be able to do in order to have a fulfilling life. For example, freedom of thought is a core capability; a social regime that prevents freedom of thought among citizens is limiting a capability necessary for flourishing life, and so the development of the country's people is stunted (Robeyns and Byskov 2023).

The approach, much like the subject matter on which it focuses, is broad. Different practitioners may have developed different methodologies for measuring and assessing capabilities. For example, some may use surveys to assess how much "access" participants have to a given capability; others may supplement survey results with census data or other forms of quantitative data (Chiappero-Martinetti 2023).

The Capabilities Approach (CA) is not specifically designed for AI implementation. The CA has been used primarily as a policy analysis tool. There also exists rich and abundant scholarship refining and critiquing it in an effort to specify its application to new contexts, such as disabilities studies (Harnacke 2013). Thus, the CA offers a family of methodological approaches all working under the general framework posited by Nussbaum and Sen. Practitioners and theorists working in development economics may have differing views on the more precise points of application of the theory, but all generally agree that one can determine (broadly) how well off people are by assessing their ability to express certain capabilities.

This approach is fruitful and diverse and will be a critical component of the methods we develop later in this document. For now, the primary concern with the CA is that, because it has been developed for assessing human development at a relatively coarse-grained level of entire countries, or at least large groups of people (e.g. a state within a country), its standard form is not ready for application in more specific domains. Nor is it designed—nor has it been applied, as

far as we know—to evaluate the deployment of artificial intelligence. Its core benefits are that it is normatively ecumenical and that its pluralism accommodates the diverse features of human experience that any impact evaluation should acknowledge.

The method we develop later on will begin from the CA, tailoring it to our specific purpose, while capitalizing on the benefits noted above.

Nussbaum and Sen provide a compelling normative argument for their view. But this is only the first step; questions remain about the ways that organizations and governments might actually implement their theoretical framework. Fortunately, there is a wealth of work in this domain. Many researchers have begun to consider the most effective ways that one might *quantitatively measure the qualitative changes* in the capabilities this approach is concerned with (message 2023). Methodologies vary, and include statistical methods native to economics and political science, as well as elements of decision theory and philosophical ethics.¹⁷ Such methods will be useful reference points for us later in the development of our own approach to metrics for impact quantification.

Legal damages (3)

Legal damages refer to our justice system’s compensation provided to those who have suffered harm or loss due to a breach of duty or violation of a right. Non-economic damages are the types of damages that are not quantified by medical bills or lost earning capacity, and arise out of bodily harm. Examples include PTSD, loss of enjoyment of life, and loss of education or opportunity (OKC Injury Lawyer 2022). In order to calculate a reasonable amount of compensation for mental distress or suffering, there are many factors that are commonly considered: e.g. the duration and severity of a person’s injuries, the degree to which the injuries affected the person’s day-to-day life, and what constitutes a full recovery from the injury or loss (Enjuris, n.d.). Because lawyers and judges are involved in attaching concrete values to nebulous harms, this project could learn from the common methods for calculating legal damages.

The two most common methods for calculating non-economic damages are the Per Diem Method and the Multiplier Method. The Per Diem Method analyzes the pain for each day of the victim’s remaining life. A reasonable dollar amount is decided and paid for each day from the time of the harm until the victim reaches maximum medical improvement. A daily compensation is typically a “reasonable sum,” usually \$100 per day or based on the minimum wage in the relevant jurisdiction (Sterling 2020). The Multiplier Method takes the

¹⁷ See Chiappero-Martinetti and Roche (2009) for an outstanding survey of these approaches.

past, present, and future medical bills the victim can expect to pay and multiplies them by some number between 1.5 and 5, which is dependent on the severity of the injury. This method is commonly used by insurance companies in deciding a settlement amount (FindLaw 2018). For example, if a man gets into a car accident that causes painful injuries and requires long-term recovery, and he is completely exempt from causing the crash, then a multiplier of 4 or 5 may be appropriate. If the medical bills from his injury amounted to \$15,000, then he may receive a settlement between \$60,000 ($\$15,000 \times 4$) and \$75,000 ($\$15,000 \times 5$) (Matthews 2021). Still, here, we see that the best that common practice can provide is a rough, reasonable, ordinal ranking of badness between 1.5 and 5.¹⁸

Political investments and priority-setting (4)

Governments are constantly tasked with distributing funds in order to, broadly, maximize public wellbeing. The study of political investments and the importation of political investment evaluation frameworks was a natural place to look. However, in general, strong evidence suggests policymakers are primarily influenced by *political considerations* when making decisions about public investment. Political cycles tend to affect economic decision-making most strongly because they give politicians incentive to stimulate the economy and increase employment before elections (Nordhaus 1975). A study from the International Monetary Fund (IMF) demonstrates that public investment increases at a rate of 2 percent of GDP in the 24-36 months before an election and grows at a negative rate starting 12 months after an election (Abdul Abiad et al. n.d.). Data from the Comparative Agenda Project reveals that policies that address immediate needs tend to be more popular with voters which motivates politi-

18 An example from the technology industry can add further detail here. *Scola v. Facebook* was a class action lawsuit filed on behalf of content moderators who developed psychological trauma or PTSD after viewing disturbing content as part of their day-to-day job at Facebook—e.g. child sexual abuse, rape, torture, bestiality, beheadings, suicide, and murder (Joseph Saveri Law Firm, n.d.). In July 2021, a settlement was delivered for the moderators' workplace litigation that brought a \$52 million fund for mental health treatment for a 14,000 member class (Wiessner 2021). All class members received a single payment of \$1,000 that may be used for medical diagnostic screenings ("*Scola v. Facebook* FAQs," n.d.). A moderator who was diagnosed with a mental health condition was eligible for additional compensation, with a maximum amount of compensation of \$50,000. The distribution of the settlement followed a methodology similar to the Multiplier Method, where the severity of mental harm and the length of recovery from such harm was taken into account. For example, a class member with a diagnosis of PTSD could be eligible for an additional \$3,000. Based on the strength of causal connection between content moderating and psychological harm, this class member could also be placed into a group where they receive up to 12x that amount (e.g. $\$1,000 + \$3,000 \times 12 = \$48,000$). Thus, plaintiffs received varying compensation based on the severity of mental anguish experienced after their time content moderating at Facebook.

cians to include short term appeal to voters in their investment considerations as opposed to long-term net positive impact (Kraft 2017). The four political factors that generally govern public investment decision making are electoral opportunities perceived by politicians, ideology of various political parties, size and agreeability of government, and quality of budgeting institutions within a legislature and do not include considerations of human well-being (Gupta, Liu, and Mulas-Granados 2015). None of these, that is to say, are tied closely to objective and neutral estimates for the aggregate impact of investment decisions.

The primary return that politicians are interested in is political capital, usually measured in the electoral success that results from the public investments they make (Bertelli and John 2013). Due to the fact that politicians usually make investment decisions based on perceived electoral payoff and not on good faith estimates of human impact, attempts to quantify electoral success are difficult to find. Furthermore, concrete evaluation criteria are also subjective because they require weights whose determination is itself subject to politics (Highsmith 2015).

Quality-Adjusted Life Years (QALYs) (5)

Quality-Adjusted Life Years (QALYs) are often used in the UK in medical and insurance settings to evaluate and prioritize courses of treatment in healthcare. QALYs are the product of two morally relevant features of a decision: the change in quality of life produced by an intervention multiplied by the change in the duration of life lived. The more QALYs an intervention has to offer, the better it is. (Note that this is an operationalization of one species of utilitarian philosophy.) If, for example, an organ transplant is more likely to produce a higher QALY score than continued medication, then it is the preferred intervention.

While they are not currently in use in any AI impact assessment contexts, it's easy to see how QALYs might be adapted to the task: once an application is deployed, simply measure the amount of QALYs produced by positively affecting the duration or quality of life—or both—of those affected. Those programs that produce more QALYs are better. Note, as well, that QALYs can be aggregated across individuals to assess outcomes and, as in classical utilitarianism, the distribution of QALYs among affected individuals is irrelevant (Whitehead and Ali 2010; Prieto and Sacristán 2003).

QALYs work well in a medical context where the desired outcome is reasonably straightforward: patient health and recovery. In a medical context, a health score of 0 denotes patient death, and 1, perfect health. This is then multiplied by years. A patient who evaluates their wellbeing at around 80% of ideal, for example, would multiply 10 years of life 0.8 to generate a QALY score of 8. This measure of 8 represents 10 life-years, whose quality is merely 0.8. The *quali-*

ty-adjusted total, then, is 8 life-years. There are two categories of concern for QALY application in AI assessment: problems of technical measurement and fundamental ethical concerns (Prieto and Sacristán 2003).

It may be reasonable to apply QALYS, including fractions of a year's worth of life, when the object being measured is one-dimensional, like *distance from peak bodily health*. But this becomes more difficult when the good being measured is replaced with others in other domains. For example, suppose a firm wishes to measure the impact of a model that determines sentencing for convicted criminals. It's unclear what the alternative to the *health* input to the *quality(-years)* expression ought to be. This produces a serious problem for AI impact qualification as the relevant objects of measure will vary significantly across domains—and AI promises to be (and already is) implemented across vastly different domains of human life. The capabilities approach we endorse, instead, includes capabilities that are not so clearly amenable to fractional approximation.

Even if such technical issues could be resolved in principle, others have pointed out that QALYS—and most renderings of utilitarian ethics—can produce seriously concerning outcomes. Even within medical contexts, QALY tradeoffs may guarantee the system provides helpful interventions to more white patients than those who belong to historically marginalized groups since (in many countries), whites have higher default health expectations anyway. That is, interventions may tend to benefit whites disproportionately, other things being equal, since they can generally avail themselves of higher-quality medical care. On a larger scale, there might be a public-health-adjusted QALY measure that maximizes QALYS for certain populations over others because the latter are already so badly off that it's "better" for the program to simply "cut its losses" and improve health outcomes for the already-healthier populations. Put another way, QALYS are *essentially indifferent* to concerns about distributive justice. Most, we suggest, will find this unacceptable. Of course, there are mathematical measures of inequality, as well, but evaluating these would add an additional level of technical complexity and theoretical baggage.

Effective altruism movement (6)

Non-profit evaluators such as Charity Navigator (CN) and GiveWell (GW) attempt to rate or rank charitable organizations based on how effective or trustworthy they are in order to give donors an idea of how to maximize the positive impact of their contribution. Once again, we are faced with the task of quantifying initiatives as diverse as literacy programs, vitamin A supplements, and mosquito nets.

Charity Navigator's rating system focuses on metrics regarding a nonprofit organization's finances and opacity rather than human impact. In order for a

charity to be rated by CN, they have to meet certain criteria regarding IRS tax status, revenue, length of operations, location, public support, fundraising expenses, and administrative expenses (Charity Navigator, n.d.). GiveWell's methodology for ranking charities, on the other hand, is more concerned with comparing charities based on human impact. A charity must serve a "priority program" that uses evidence-based approaches to help the global poor. GiveWell evaluates charities using four criteria: evidence of effectiveness, cost-effectiveness, transparency, and room for more funding. Charities must be able to offer empirical evidence that their efforts have resulted in improved life outcomes, such as providing mosquito nets that lead to lower rates of malaria in an affected region (GiveWell, n.d.). The manner in which GiveWell gauges cost-effectiveness of a charity is primarily determined by the extent to which the organization saves or improves lives for as little money as possible. GiveWell looks at metrics such as the cost per life or life-year changed (death averted, year of additional income, etc.), life saved per dollar, or proportional increase in income per dollar donated (GiveWell, n.d.).

While often an effective approach for donating money to worthwhile charity organizations, EA would likely have several drawbacks as a method for AI impact quantification. First, EAers (as they are known) don't have a set formula or algorithm that generates a single answer for which charities are best. They can provide considerations or general principles for where to donate: these include marginal impact (counterfactuals), neglectedness, and outputs in terms of human wellbeing (similar to QALYs above). Some of these (like "neglectedness") are relevant when considering optimal donations to charities, but less clearly so for other domains where AI might be implemented.

Additionally, EAers are prone to "quick and dirty" comparisons because, as outlined elsewhere in this report, they too are beset by the challenge of comparing apparently incommensurable benefits, e.g. the value of donating to improve education vs. the value of donating to improve public health. They nonetheless think you can make at least orders of magnitude comparisons relatively reliably. For example, aiding education in one particular school district, while still producing some good, is not nearly as good as donating money to a charity that provides mosquito nets to populations with a high risk of malaria. So, they reason, it's almost always better to do the latter than the former. In practice, EAers often end up giving lists of charities that are qualified or meet a minimum threshold rather than saying that some unique charity dominates all the others, i.e. that it is superior in every respect.

There are two practical problems with EA. First, because EA is concerned with maximizing good, there are some EAers who endorse Longtermism (see concerns with "Longtermism and AGI" above). This approach is too broad to be helpful in most of the smaller-scale situations where an AI might be imple-

mented. The second concern is related to that of QALYs above. While an AI's impact on general wellbeing is essential to any impact assessment, there are many other goods that are worth assessing in their own right, and which EA methodology does not have the tools to do well.

Compensating the wrongfully imprisoned (7)

A similar, adjacent task of the legal system is compensating the wrongfully imprisoned. According to the Innocence Project, individuals proven to have been wrongfully convicted spend, on average, more than 14 years behind bars (Innocence Project, n.d.). Life after prison presents profound challenges in establishing a professional status, housing, transportation, health services, and insurance. Incarceration can also be traumatic and psychologically debilitating. Some research suggests that people in prison “experience mental deterioration and apathy, endure personality changes, and become uncertain about their identities” (Kregg 2016). Time away from friends and family also compounds economic hardships and the ability to re-enter society.

Appropriately compensating former prisoners for a suite of various harms presents serious theoretical and practical challenges, and the law varies widely across jurisdictions.¹⁹ In California, wrongfully convicted felons are compensated \$140 per day ([ca.gov](https://www.ca.gov/)); Missouri provides exonerees \$50 per day. Wisconsin caps the total amount of compensation at \$25,000 and Maine has a cap of \$300,000 (Legner and Arndt 2022). Texas has taken steps to provide non-monetary compensation for exonerees through the Tim Cole Act: including medical, dental, and psychological care, as well as assistance in completing the necessary paperwork for such entitlements (Legner 2022).

Quantitative human rights (8)

Rights are a common and robust way to frame our moral obligations. Cases where human rights are respected are good, and cases where they are violated are bad and to be avoided. Some theories of rights require, in fact, that rights function as “constraints” on pursuing the good, which is to say that rights violations are never acceptable, regardless of how much good could be done.

¹⁹ Currently, the federal government, the District of Columbia, and 38 states have established compensation statutes. Federal compensation law provides \$50,000 per year of wrongful incarceration. President George W. Bush endorsed Congress's recommended amount of up to \$50,000 per year, with an additional \$50,000 per year spent on death row. Adjusted for inflation, this amount is \$63,000 (Innocence Project). Still, the presence of a compensation statute in the 38 states with them does not necessarily mean that the compensation will be received and filing for compensation can often take years (Legner 2022).

Recently, some legal, human development, and other scholars have attempted to quantify the assessment of respect for rights. Such attempts use statistical methods and “big data” to answer questions like: How often do physical rights abuses happen? How often are speech rights abused? How often do rights to healthcare get violated? These methods then assess the impact of, say, a policy by determining the total number and severity of rights violations that might occur as a result.

This method provides interesting insights into the way that something like *rights violations* can be aggregated and measured to better bolster arguments against policy interventions. Say, for example, if some policy results in a much higher frequency of free speech rights violations, one might argue that it is worse, regardless of how much economic activity it might be expected to generate. There may be cases where AI impact can be measured using such methods. For example, criminal sentencing algorithms may result in a higher or lower frequency of rights violations than human sentencing decisions. But while such assessments may be a crucial element of AI impact assessment in some contexts, there are serious concerns with a more general application of this method to AI impact quantification.

One weakness of this methodology is that they lack a normative foundation of which rights are worth protecting. Such methods must rely on *some* normative view to determine which rights are the ones worth caring about and when. Without that normative view, which invites its own tangled discussion, an injunction to care about rights, on its own, is not very helpful. While a standard set of human rights—like those contained in the International Covenant on Civil and Political Rights (ICCPR) or International Covenant on Economic, Social and Cultural Rights (ICESCR)—does provide some normative footing, this is far too coarse grained to be useful in assessing something like, say, hospital triage algorithms.²⁰

Philosophical Challenges

Understandably, any attempt at creating a metric for evaluating policy and AI impacts that is robust, accurate, and operationalizable will face challenges.²¹

20 For examples, see Gibney and Haschke (2020).

21 Perhaps the most confounding problem plaguing every attempt at impact quantification is the fact that there are *unknown unknowns*, in the eternal words of Donald Rumsfeld (2002). We may be able to identify some issues ahead of time. These are the known knowns. There are, further, variables and influences that we know of but do not know or cannot account for. These are the known unknowns. But there are likely many possible issues that we simply

Some of the most frustrating challenges to developing such metrics are *philosophical*. These challenges manifest at the very foundation of an approach: if a given attempt to measure outcomes is philosophically flawed, then not only will practical problems likely follow, such problems will be “baked into” its use. As an example, consider the willingness-to-pay metric discussed above. If such a metric is reliant on the premise that *the goodness of all outcomes are monetarily evaluable*, and that premise is false, then all attempts to create evaluative frameworks that rely on this premise will fall prey to the same problems. Just as significantly, the willingness-to-pay metric relies on a proxy that systematically distorts the importance of outcomes if, e.g., the rich are willing to pay more than the poor to avoid some outcome, then the priorities and preferences of the rich are systematically inflated in importance.

In this section, we outline several of the philosophical problems that beset many of the attempts cased above as candidates for impact quantification. There’s hard work that must take place at different ends of this issue: at the normative end, in terms of deciding how to represent the concept of wellbeing, what the relevant tradeoffs are, and what the absolute red lines are; and at the practical end when designing an operationalizable metric that can be quickly and reliably applied while meeting the requirements outlined at the normative stage. The following items are problems to be avoided during the normative phase.

Measurability bias

In the search for a unit of evaluation that is quantifiable, there is the temptation to push out other variables that are harder (or perhaps impossible) to measure. Household income or the spread of disease are easy to measure; like political freedom and equality are very difficult to measure. The trouble is that, often, things that are difficult to measure are quite important. Less measurable things are often structural, comparative, and are beset by collective action problems. This is opposed to features of a problem that can be measured or affected by a single agent, like income or frequency of mechanical failure. There may be, then, a tendency to select measurements that are easier to track, rather than those that are actually the most indicative of the impacts of a given AI system. The number of users actively engaged on a social media platform is easy to

cannot foresee—nor can we foresee that we cannot foresee them. Before Donald Rumsfeld, this same frustration gave rise to one horn of the Collingridge Dilemma (Genus and Stirling 2018): (1) while a technology is in development, it is easiest to make modifications in anticipation of its effects, but its effects are least clear. (2) These effects become clear only after the technology has been deployed but at which point, ironically, it is hardest to make changes to ameliorate any negative impacts.

measure, and may indicate excitement for a platform; the social malaise, sense of isolation, or political disruptions such platforms engender is not.

Moreover, and perhaps conclusively, the worst off populations around the world often pass over easily quantified metrics and point to more nebulous concerns when they are asked to report their greatest concern. See Lechterman on this point:

As Monique Deveaux reports, persons facing severe want tend not to point to physical pain or material discomfort as their chief concerns. Rather, they describe overriding senses of powerlessness, shame, and humiliation, as well as resentment towards the arbitrary commands of local authorities. (2020, 102)²²

Any effort concerned to faithfully represent the concerns of the worst off in society ought to supply, therefore, some guidance on measuring impacts such as powerlessness, shame, humiliation, and so on. These are much harder to quantify, of course, than more popular metrics like gross domestic product (GDP) per capita.

Box-ticking and ‘Goodharted’ metrics

Another category of concern is box-ticking. Box-ticking occurs when the implementation of a method of impact evaluation encourages administrators to satisfy certain conditions without considering those conditions as genuine ethical concerns. There are two species of this problem: (1) certain ethical frameworks, for example, highly legalistic ones, may themselves encourage administrators to build and assess AI systems such that they are compliant, even if compliance is only loosely related to what is ethical²³ and (2) no matter the system, there is always the concern that administrators will simply choose to treat it as a box-ticking exercise, hitting targets on, say, equity or frequency metrics without concern for the ethical content of the outcomes. This becomes problematic in those inevitable cases where developers are faced with a challenge that is not clearly and directly accounted for in the metrics they have ready-to-hand.

22 See, further, the Deveaux article that Lechterman references, and the original Voices of the Poor survey from which these insights derive (Deveaux 2015; Narayan and Petesch 2002).

23 This is sometimes known as Goodhart’s law: once a metric becomes a target for optimization, it ceases to be a useful metric (Goodhart 1975; Strathern 1997; Ravetz 1971). This epitomizes the concern that metrics can be “gamed” or reveal perverse behavior including reward hacking (see elsewhere in this report). But as Cathy O’Neil has argued persuasively, if a metric is well-crafted, and if it successfully captures what we want it to capture, then we should want people to optimize their behavior toward those metrics (O’Neil 2016).

Species (2) may be a matter better considered in training materials or hiring criteria for AI administrators, compliance professionals, or project managers. However, (1) is a matter of concern for our work here. Ensuring that methods of evaluating impact are not reducible to a box-ticking exercise should be a design priority for any metric evaluation system. This type of problem has several manifestations; the next example, monism, is one such manifestation.

Monism

Another family of philosophical critiques is related to approaches that are tethered to a form of monism: this is the view that there is only one value which all other values are reducible to. This view is *theory-bound* because in order to justify the claim that all other values are reducible to the value of choice, one must endorse a particular normative theory at the exclusion of others.

For example, quality-adjusted life-years (QALYs) are an operationalized version of a broadly utilitarian view that all the various dimensions of a person's flourishing or languishing can be measured in terms of the *quality and duration of their life*. Similarly, and much more implausibly, willingness-to-pay reduces claims about valuing different goods to an assessment of the amount someone is willing to pay to secure those goods.

There are at least three sorts of worries about monist approaches. First, monism is itself philosophically dubious; there are strong reasons to think that values are not reducible to a single overarching value. There are plausible arguments for competing theories of the good life, and 2,500 years of moral philosophy has not determined a winner (yet). Moreover, endorsing one vision of the good life at the expense of others risks being paternalistic or culturally imperialist. A pluralist approach open to multiple instantiations and sensitive to cultural contexts is more acceptable for these reasons.

Second, even if monism were true, determining precisely which variable was the right one is challenging. Utilitarians may point to wellbeing or happiness, but operationalizing this is tricky, and many utilitarians may even disagree with the simplicity of QALYs as a fair rendering of the central value they are interested in. Similarly, monism fundamentally neglects the *distribution* of goods among people, which many people take to be intrinsically valuable—i.e. its own source of justice or injustice. Classical utilitarianism, for example, regards two distributions as equivalent, as long as they sum to the same total amount of happiness, and regardless of how those benefits are distributed among society.

This last point is not only philosophically suspect, it also creates practical problems. Reward hacking is a problem wherein a program realizes some specified goal that is ultimately not aligned with its administrators' intended outcome. It is much easier for this to happen in a system which has a proxy variable

that it is programmed to maximize, and which may end up trampling other important values in the process (Skalse et al. 2022). For example, a system built to triage hospital patients to optimize health outcomes may choose to have only those with the most minor injuries and illnesses seen by medical staff, thus improving the rate of successful recovery, but at the expense of the value that is actually important: the general wellbeing of all those who come to the hospital seeking it. Accommodating multiple values in our reward function is one way of defending against the possibility of reward hacking.

Theoretical ungroundedness

Consider examples from above like Value Sensitive Design or Model Cards. Approaches like this purport to incorporate some element of value into the design or implementation of AI systems. And to do this they rely on some basic sense of what humans value (perhaps something like equity, justice, or diversity).

The concern with approaches like these is, in a sense, the opposite of the monist's problem. While monists rely entirely on the correctness of their philosophical views, other approaches attempt to implement values in the design of AI without a clear theoretical grounding for those values. That is, though a method of evaluation may claim to value equity, there may not be a framework that explains why this is the case, or to what degree it is valued.

The latter point is especially crucial in cases where AI systems and their administrators must make decisions between two variables. For example, suppose that overall wellbeing can be increased at the cost of equity. Should an administrator evaluating the impact of an AI system determine that this choice was morally worthwhile? Without a clear articulation of the moral theory supporting the values used to assess an AI's impact, there is not really an answer. This presents both a theoretical implausibility (presumably *some* choice has stronger reasons in its favor) and a practical conundrum. Without an overarching theory which can adjudicate disagreements, both the theorist's work and the computer scientist's work are threatened.

While many of the value-incorporating tools discussed above contain helpful mechanical features for measuring outcomes of different kinds, *evaluating* those quantitative results is not possible without some sort of operational value framework. Thus, while several of the metrics outlined above may be helpful features of an AI assessment process—producing some sort of quantitative output to measure a value—they do not themselves determine the values.

Requirements for a Successful Approach

Moving forward with these philosophical challenges in mind gives us a clear set of conditions for a successful approach to impact quantification. The most appropriate metrics for our project must meet the following criteria.

- **Metrics must track meaningful features of wellbeing.** As noted above, there is a danger that practitioners assessing impact might use the most easily accessible or prevalent data at the expense of more meaningful ways of assessing the wellbeing of people affected. For example, an easily measurable metric like GDP per capita might be selected over something harder to measure, like a person's political power. A successful approach will quantitatively assess features of people's lives that more *accurately* and *holistically* indicate wellbeing, not just those that are easy to measure.
- **Metrics must include sufficient nuance to avoid box-ticking.** Legalistic or crudely quantitative measures of wellbeing may incentivize practitioners to merely tick-off boxes when assessing an algorithm's impact. A successful approach will include sufficient nuance such that the result of an impact evaluation represents the *actual* wellbeing of those affected, not merely a proxy. This means delivering a full assessment of the relevant features of wellbeing in an assessment, and avoiding schemes with minimum benchmarks. In practice, those assessing impact should also be incentivized to actually care about the accuracy and quality of their evaluation, which requires a greater investment in organizational ethics and culture.
- **Metrics must not be over-reliant on one measure of goodness.** Monistic approaches (those that only take into account one desideratum, like QALYs or willingness-to-pay) can easily be reward-hacked at the expense of other outcomes we care about (like justice or community). Thus, a successful approach to impact quantification will include a pluralistic understanding of human flourishing, such that systems by design reinforce the aspects of human life that make it valuable and worth living. This means often including several dimensions of wellbeing, rather than one, and several metrics for each dimension of flourishing.
- **Metrics must be rigorously grounded in normative theory.** Lastly, any successful impact quantification metric must be grounded in a defensible conception of *the good*. Without a clear conception of the ethical considerations that underlie an impact quantification method, there is no way to determine whether an outcome is good or bad, or in what way it might need to be changed. Thus, any successful method must include a rigorous explanation of what moral theory (or at least minimal axiological commitments) grounds its evaluations. This requirement is also critical to

building transparent systems of impact evaluation, which are open to public debate and refinement.

The Universality of the Capabilities Approach

Any account of *the good*—a theory that explains what it means for humans to have a good life—will run up against a common objection: there are many different ways that people can choose to make a good life. Indeed, this is a foundational principle of Western democracies. We are committed to the belief that different people of different backgrounds may come together and, in tolerance and respect, choose what sort of life is best for each of them without interfering in others' lives. A stronger version of this view is *relativism*, the position that all moral goodness is ultimately determined only by cultural situation or custom.

So, the thought that it is possible to develop a universally applicable scorecard for the measurement of human impacts across domains and cultures may seem, to some, as arrogant, illiberal, or impossible. However, this is decidedly not the case. We hold that it is possible to develop a universal account of human wellbeing that remains sensitive to the diversity inherent in human culture and across different domains.

Importantly, the capabilities outlined in “An Ecumenical Theory of Human Flourishing” above are *non-relative*. That is, though the way that a culture or area of human activity may promote or intersect with these capabilities may vary, the fact that human well-being relies on the maintenance of those capabilities (regardless of cultural situation) means that they are themselves non-relative. So, this approach provides a sturdy theoretical foundation for an operationalizable method for measuring human impact that embodies two theoretical virtues: context-sensitivity and non-relativity.

The capabilities listed above are plural (they are not reducible to each other) and multiply-realizable (many different methods of organizing society can effectively nourish those capabilities). These features of the view secure many of the concerns that motivate the relativist: that diversity in human culture should be respected, and that not all goods are reducible to one governing good. However, they also avoid the theoretical problems relativism is open to: an inability to morally assess any cultural practice, tolerance for intolerance, and other intuitively implausible outcomes.

Martha Nussbaum develops an argument for this sort of non-relative account of virtue in her chapter, “Non-Relative Virtues: An Aristotelian Approach” (2000). Her argument makes two important moves, which we take to support the view that we develop in this report. First, she notes that virtues

tend to be plural, and acknowledges that in many cases, a plural account of *the good* seems more plausible at first pass than the utilitarian-friendly QALYs or Kantian-inspired human rights. Arguably, the goods a culture endorses (e.g. generosity, equality among genders, freedom of exchange) may vary significantly from culture to culture, and so an ethical theory that focuses on such a plurality of goods is predisposed to be relativistic—or so this line of thinking might go.

The trouble with a fully-fledged relativistic approach is that it renders assessment nearly impossible. If the condition for the goodness or badness of some policy is whether it accords with local norms, we come upon implausible results. An AI that reproduces systemic gender biases, for example, may be said to be reproducing local norms. But this is an unacceptable result. Hardly any moral philosophers accept moral relativism for this and similar reasons.

Some sort of universally applicable set of measurements for human impacts and wellbeing must be possible. However, the fact remains that human cultures—and the domains where people work and AI will be implemented—are diverse. And so theoretical frameworks (like utilitarianism or Kantianism) that are less sensitive to this variation may be unfit to serve within the myriad domains where AI will operate (see the challenges with monistic approaches in the section “[Philosophical Challenges](#)” above). Offering QALY scores to newsroom editors using natural language processors, for example, may seem like an irrelevant intrusion. An ideal metric for assessment must take into account the apparent plurality of existing values while maintaining that *plurality does not entail relativism*.

The second move Nussbaum makes resolves this tension. She argues,

The fact that a good and virtuous decision is context-sensitive does not imply that it is right only *relative to*, or *inside*, a limited context, any more than the fact that a good navigational judgment is sensitive to particular weather conditions shows that it is correct only in a local or relational sense. It is right absolutely, objectively, anywhere in the human world, to attend to the particular features of one’s context; and the person who so attends and who chooses accordingly is making, according to Aristotle, the humanly correct decision, period. If another situation should ever arise with all the same ethically relevant features, including contextual features, the same decision would again be absolutely right. (2000, 257)

The key to this approach is that the many particular customs and mores apparent across human culture can be subsumed under a broader category that cuts across cultural variation. Across cultures, there exist ways of managing

goods that occur in different spheres of human experience: food, family, societal organization, danger and death, sex, and so on. This non-relative approach, then, seeks to promote well-being by promoting different human *capabilities* across these spheres. The aim is that, by nurturing these base capabilities, the ability to access those goods increases, keeping in mind whatever local mores might govern a given culture or domain. In this way, the approach is context-sensitive while still maintaining an absolute conception of moral goodness.

Some Mature Quantitative Impact Metrics

Dimensions of Wellbeing	Examples of Relevant Metrics
A full life	The most straightforward measure of a “full life,” as conceived by Nussbaum and Sen, is simply life expectancy—for example, calculated based on the probability of death at any given age (Canudas-Romo 2010). Availability of healthcare may be another measure; ML models could conceivably facilitate access to affordable medical care (Niëns and Brouwer 2013).
Bodily health and integrity	A rough metric for bodily health, as independent from life expectancy, would include multiple indicators (weight, height, food access, etc.) (Furlong et al. 2016). Measuring children’s weight for their age is a common way of assessing whether they are receiving sufficient nutrition (Di Tommaso 2007). Affordability of medicine may also prove a telling metric (Niëns and Brouwer 2013). Sen suggests using the ratio of hospital inpatient admissions to hospital deaths as an indicator of the quality of medical care in a country (1985). Others might include access to a healthy nutritional environment available to a population (Glanz et al. 2005). As with a full life, access to healthcare might also be used to measure bodily health and integrity (Wendt 2009).

Dimensions of Wellbeing	Examples of Relevant Metrics
Imagination and thought	<p>The most common metric for capturing the dimension of imagination and thought is education in some form (Brandolini and D'Alessio 1998, (Chiappero-Martinetti 2023). This in turn can be measured in different ways. For example, the UNDP uses a composite metric for educational attainment, combining adult literacy (two-thirds weight) and combined primary, secondary and tertiary enrolment ratios (one-third weight) (UNDP, 1995). Another way imagination and thought might be measured is the availability of educational institutions beyond schooling: art museums, science museums, aquariums, libraries etc., as well as access to sources of creative engagement—internet access, for example (Valentín-Sívico 2022).</p>
Sensation and Emotion	<p>Mental wellbeing, including metrics such as subjective wellbeing reports are promising, easy to access, and relatively straightforwardly tied to the operation of ML models in certain contexts (Ranis, Stewart, and Samman 2006).</p> <p>Capturing the quality of a person's sensations and emotions, in a way that is independent of their subjective wellbeing report, turns out to be a challenge. One way of doing this would be to measure specific activities that contribute to emotional wellbeing, such as rates of sexual activity (Ueda et al. 2020), or level of reported loneliness (Weissbourd et al., n.d.).</p> <p>Another way of capturing the quality of a person's sensation and emotion is to look at negative indicators, such as rates of substance abuse and drug addiction and frequency of suicidal ideation (HHS, n.d.).</p>
Practical reason	<p>The UN's gender empowerment measure (GEM) is a composite of factors meant to reflect the ratio of men's and women's abilities to participate meaningfully in society. This includes income (PPP), share of jobs that are technical, administrative, or professional, and the relative share of parliamentary seats in government (UNDP 1995). Other metrics might include access to information & press freedom ("Homepage RSF" 2023), freedom of worship (Fox, 2021), and freedom of speech (Staghøj and Krishnarajan 2021).</p>

Dimensions of Wellbeing	Examples of Relevant Metrics
Affiliation	Some studies (Kawamichi et al. 2016; Chiappero-Martinetti and Roche 2009) measure frequency and quality of social interactions and their effects on wellbeing. Such points of data (frequency of social interaction; time spent with friends; closeness to family) might be helpful indicators when compared with the rate that occurs after the introduction of an AI-based technology. Other plausible metrics to assess affiliation include social, psychological, and physical contact and number of social relations (Brandolini and D'Alessio 1998).
Other species	Indicators of access to other species might include biodiversity measures (Rousseau and Van Hecke 1999), access to protected natural areas (Holland et al. 2021), and pet ownership (American Veterinary Medical Association n.d.).
Play	Direct report via questionnaire is often the easiest way to get ahold of this information, for example, asking people to estimate their free time and if they are satisfied with that overall. Other metrics include Internet, radio, television penetration, and movie attendance. Rates of participation in sports and other recreational activities may be similarly informative (Schokkaert and Van Ootegem 1990; Di Tommaso 2007), as well as work-life balance reports (Fisher and Layte 2002).
Control over one's material conditions	Several key indicators might inform whether a population has a high degree of control over material conditions: the percentage of income spent on rent, availability of housing, rates of homelessness, rate of political participation (UNDP 1995), rate of union membership, rates of unionization across different industries ("U.S. Bureau of Labor Statistics" n.d.), and percentage of income spent on rent ("Renter Cost Burdens Reach Record Levels Joint Center for Housing Studies" n.d.). Further measures of the quality of housing include people per room and access to heating, insect infestations, and safety indicated by things like exposed wires (Brandolini and D'Alessio 1998). Participation in the labor market is another common metric, including job quality and satisfaction, whether people are employed, underemployed, or have stopped looking for work (Brandolini and D'Alessio 1998).

Dimensions of Wellbeing	Examples of Relevant Metrics
Healthy natural environment	The Yale Center for Environmental Law and Policy (YCELP) at Yale University and the Center for International Earth Science Information Network (CIESIN) at Columbia University have developed a composite measure, including 40+ indicators, of a country's "environmental performance." The most relevant for our purposes would include exposure to PM2.5, NOx, lead, and other hazardous chemicals; recycling rates, and the status of fish stocks (Yale Center for Environmental Law and Policy (YCELP) at Yale University, and Center for International Earth Science Information Network (CIESIN) at Columbia University 2022). Other indicators include air quality (Murena 2004), water quality (EPA Fresh Water Quality Index), and the level of accessibility of park and outdoor spaces to people living in an area (Dai et al. 2022).

Next Steps

Indeed, a set of quantitative impact metrics that can be “plugged into” the evaluation of machine learning models is a holy grail for the industry. The work cannot be completed in the short 15-month span of this project. We have put forth a strong proposal for advancing the state of the art, but there is much more to be done to make the ends meet.

This final section serves as a roadmap to scaffold future iterations of this work. Future work could include refining any of the outputs of this current project, e.g. identifying reliable and valid quantitative impact measures (QIMs) for particular domains; piloting those QIMs in small projects; and partnering with technology vendors to understand how QIMs could be integrated into their product specifications and development processes.

Sociotechnical Analysis

Machine learning models are frequently integrated into complex sociotechnical systems that incorporate human interaction and feedback. Evaluating the impact of these models requires accounting for the human behaviors within these systems, including data input, interpretation of recommendations, and decisions to act (or not) on those recommendations. So, it is not straightforward to trace a direct, unbroken path from the model’s design and evaluation to its ultimate human impacts.

Liu (2019) serves as an illustrative example. In this instance, the authors had access to data on the ultimate consequences of different lending decisions by a bank. This data allowed them to simulate the downstream human impacts of various models. However, the path between the decisions of models and their real-world effects is not always so direct. Oftentimes, there is ample room for human judgment, especially outside of stringent environments like banking. Moreover, obtaining data on the ultimate impacts can be challenging.

In such cases, we might struggle to integrate these long-term utilities directly into the models themselves. A key question emerges: Could contextual or sociotechnical factors overshadow the marginal difference in impacts stemming from the design of models? Could there be a significantly greater return on investment, in terms of human impact, if we focused more on UI/UX factors that shape user interpretation of the system and ultimately drive their behav-

ior? (This is not to suggest a false dichotomy, of course, since both projects are worthwhile and ought to be pursued energetically.)

Other design considerations, such as the possibility of overruling algorithmic decisions, play an essential role in this scenario. However, this can also lead to unintended consequences. For example, a bail algorithm in Kentucky, which judges could override, resulted in judges disproportionately overruling decisions for black defendants, leading to higher bail amounts (Albright 2019; Covert 2020; Simonite 2019).

In general, our confidence in a model's impacts is blurred or 'fuzzed' by sociotechnically contingent factors. This downstream uncertainty is why measures of the models themselves *in isolation*, such as fairness and accuracy metrics, have gained prominence. By comparison, they offer an enviable level of accuracy and precision. Meanwhile, it has been more challenging to confidently evaluate concrete human impacts, which are at a greater causal distance from the design of the model.

However, the inherent flexibility of sociotechnical systems allows for the possibility to "push" these values elsewhere throughout the system. A compromise in one area can be offset elsewhere in the system (van de Poel 2015). This is a common approach in systems engineering. For instance, in commercial airlines, much of the concern for safety is transferred from the plane's design to the design of the systems surrounding the plane, such as air traffic control, with human behavior being strictly regulated.²⁴ Because airplanes cannot be made out of lead, finding a compensating level of safety elsewhere in the system is crucial.

Other Species of Harm

It will be crucial to consider the universe of possible harms and whether they can be captured adequately by the theory of wellbeing we propose. Some harms have risen to prominence in the public conversation around artificial intelligence, and two of these seem especially concerning and especially difficult to quantify and accommodate.

First, concerns about privacy harms are perhaps the most common worry when it comes to artificial intelligence. If privacy harms are viewed as intrinsically valuable, then it may be necessary to add a dimension of wellbeing to our above theory. However, the capabilities we already include could plausibly account for the harms of privacy violations, for example: (1) an individual's

²⁴ This example was shared by David Danks in conversation.

subjective wellbeing or sense of security are undermined by actual or suspected privacy violations (Calo 2011), and (2) metrics that measure government accountability, press freedom, religious freedom, or freedom of affiliation could serve as useful proxies. Given that privacy violations tend to give rise to specific types of harm, much of the concern about these violations could be accounted for through their effects on other capabilities that are already included.

Another important dimension to consider is representational harm, which could arise in the context of search, advertisements, and image recognition—for instance, through the propagation and perpetuation of stereotypes (Noble 2018; Carpenter 2021). Much like privacy harms, representational harms represent a newly recognized category of harm (Wang et al. 2022; Mehrabi et al. 2021; Buddemeyer, Walker, and Alikhani 2021). Representational harms might be a *sui generis* species of harm—although it also seems possible to account for them through the metrics we already include. Alternatively, we might need to include a new capability akin to non-domination, freedom from oppression, or dignity. Such a metric would capture certain kinds of harm typically associated with dimensions of race, gender expression, class, sexuality, religion, and so forth.

These additional species of harm are also relevant when considering the impacts of large language models (LLMs). These models, which are increasingly popular and influential, are implicated in a wide variety of potential harms, including breaches of privacy, issues of intellectual property, the propagation of representation bias, the dissemination of disinformation, and the potential for self-harm. This makes their evaluation much more challenging because of their domain-agnostic nature and flexibility. While some of these concerns are captured within the current measurements we put forth here, others prove more elusive. Addressing these specific dimensions would enrich our understanding and management of potential harms arising from LLMs.

Like the ongoing project of moral philosophy from which it springs, our project relies on a collated and systematized account of what human beings take to be valuable, i.e. what are the components of a life worth living. A satisfying instrument, then, would be able to accommodate the full range of ideas about the good life that reasonable, reflective people have defended—whether these harms were inaugurated by the age of AI or whether they were contemplated thousands of years ago. Understanding the complex interplay of these various types of harm could lead to more comprehensive and effective strategies for managing and mitigating the impacts of AI.

Works Cited

- Abbott, Andrew. 1991. "The Order of Professionalization: An Empirical Analysis." *Work and Occupations* 18 (4): 355–84. <https://doi.org/10.1177/0730888491018004001>.
- Abdul Abiad, Aseel Almansour, Davide Furceri, Carlos Mulas-Granados, and Petia Topalova. n.d. "Is It Time for an Infrastructure Push? The Macroeconomic Effects of Public Investment." Accessed June 24, 2023. https://www.eco.uc3m.es/temp/ppt_WEO_Ch3.pdf.
- Albright, Alex. 2019. "If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions." *Law, Economics, and Business Fellows' Discussion Paper Series* 85.
- American Veterinary Medical Association. n.d. "U.S. Pet Ownership Statistics." American Veterinary Medical Association. Accessed June 24, 2023. <https://www.avma.org/resources-tools/reports-statistics/us-pet-ownership-statistics>.
- Aristotle. 2019. *Nicomachean Ethics*. Translated by Terence Irwin. Third edition. Indianapolis: Hackett Publishing Company, Inc.
- Beitz, Charles R. 2009. *The Idea of Human Rights*. Oxford: Oxford university press.
- Bertelli, Anthony, and Peter John. 2013. *Public Policy Investment: Priority-Setting and Conditional Representation in British Statecraft*. Oxford Academic. <https://academic.oup.com/book/12209>.
- Brandolini, Andrea, and Giovanni D'Alessio. 1998. "Measuring Well-Being in the Functioning Space." *General Conference of The International Association for Research in Income and Wealth, Cracow, Poland*, August.
- Brockman, Ben, Hersh, Skye, Gosselink, Brigitte Hoyer, Maganza, Florian, and Berman, Micah. 2021. "Investing in AI for Good." <https://doi.org/10.48558/R12S-MP19>.
- Buddemeyer, Amanda, Erin Walker, and Malihe Alikhani. 2021. "Words of Wisdom: Representational Harms in Learning From AI Communication." arXiv. <https://doi.org/10.48550/arXiv.2111.08581>.
- Calo, Ryan. 2011. "The Boundaries of Privacy Harm." *Indiana Law Journal* 86: 1131.

- Canudas-Romo, Vladimir. 2010. "Three Measures of Longevity: Time Trends and Record Values." *Demography* 47 (2): 299–312. <https://doi.org/10.1353/dem.o.0098>.
- Carpenter, Julia. 2021. "Google's Algorithm Shows Prestigious Job Ads to Men, but Not to Women. Here's Why That Should Worry You." *Washington Post*, October 26, 2021. <https://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you/>.
- Cath, Yuri. 2016. "Reflective Equilibrium." In *The Oxford Handbook of Philosophical Methodology*. Oxford University Press.
- Charity Navigator. n.d. "Charity Navigator's Methodology." <https://www.charitynavigator.org/index.cfm?bay=content.view&cpid=5593>.
- Chiappero-Martinetti, Enrica. 2023. "A Multidimensional Assessment of Well-Being Based on Sen's Functioning Approach."
- Chiappero-Martinetti, Enrica, and José Manuel Roche. 2009. "Operationalization of the Capability Approach, from Theory to Practice: A Review of Techniques and Empirical Applications." *Debating Global Society: Reach and Limits of the Capability Approach*, 157–203.
- Covert, Bryce. 2020. "How a Bail Reform Tool Failed to Curb Mass Incarceration." *The Intercept*. <https://theintercept.com/2020/07/12/risk-assessment-tools-bail-reform/>.
- Craigie, Jillian. 2011. "Thinking and Feeling: Moral Deliberation in a Dual-Process Framework." *Philosophical Psychology* 24 (1): 53–71. <https://doi.org/10.1080/09515089.2010.533262>.
- Dai, Weiwei, Suyang Yuan, Yangyang Liu, Dan Peng, and Shaofei Niu. 2022. "Measuring Equality in Access to Urban Parks: A Big Data Analysis from Chengdu." *Frontiers in Public Health* 10 (October): 1022666. <https://doi.org/10.3389/fpubh.2022.1022666>.
- Daniels, Norman. 2003. "Reflective Equilibrium," April. <https://plato.stanford.edu/ENTRIES/reflective-equilibrium/>.
- Deloitte. 2018. "AI and Risk Management." <https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/financial-services/deloitte-uk-ai-and-risk-management.pdf>.
- Deuze, Mark. 2005. "What Is Journalism?: Professional Identity and Ideology of Journalists Reconsidered." *Journalism* 6 (4): 442–64. <https://doi.org/10.1177/1464884905056815>.
- Deveaux, Monique. 2015. "The Global Poor as Agents of Justice." *Journal of Moral Philosophy* 12 (2): 125–50. <https://doi.org/10.1163/17455243-4681029>.

- Di Tommaso, Maria Laura. 2007. "Children Capabilities: A Structural Equation Model for India." *The Journal of Socio-Economics, The Capabilities Approach*, 36 (3): 436–50. <https://doi.org/10.1016/j.socec.2006.12.006>.
- Diener, Ed. 1994. "Assessing Subjective Well-Being: Progress and Opportunities." *Social Indicators Research* 31 (2): 103–57. <https://doi.org/10.1007/BF01207052>.
- Diener, Ed, and Katherine Ryan. 2009. "Subjective Well-Being : A General Overview." *South African Journal of Psychology* 39 (4): 391–406. <https://doi.org/10.10520/EJC98561>.
- Enjuris. n.d. "California Pain & Suffering Damages: Calculate Emotional Distress." <https://www.enjuris.com/california/pain-and-suffering-damages.html>.
- Evans, Jonathan St. B. T. 2008. "Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition." *Annual Review of Psychology* 59 (1): 255–78. <https://doi.org/10.1146/annurev.psych.59.103006.093629>.
- Facebook. 2020. "How Facebook Uses Super-Efficient AI Models to Detect Hate Speech." Meta AI. November 19, 2020. <https://ai.facebook.com/blog/how-facebook-uses-super-efficient-ai-models-to-detect-hate-speech/>.
- FindLaw. 2018. "What Is a 'Pain and Suffering Multiplier'?" November 30, 2018. <https://www.findlaw.com/injury/car-accidents/what-is-a-pain-and-suffering-multiplier.html>.
- Fisher, Kimberly, and Richard Layte. 2002. "Measuring Work-Life Balance and Degrees of Sociability: A Focus on the Value of Time Use Data in the Assessment of Quality of Life." 32. EPAG Working Papers. European Panel Analysis Group (EPAG). <https://www.esri.ie/system/files/media/file-uploads/2002-10/OPEA21.pdf>.
- Forsyth, Patrick B., and Thomas J. Danisiewicz. 1985. "Toward a Theory of Professionalization." *Work and Occupations* 12 (1): 59–76. <https://doi.org/10.1177/0730888485012001004>.
- Fox, Jonathan. 2021. "What Is Religious Freedom and Who Has It?" *Social Compass* 68 (3).
- Friedman, Batya. 1996. "Value-Sensitive Design." *Interactions* 3 (6): 16–23. <https://doi.org/10.1145/242485.242493>.
- Friedman, Batya, David G. Hendry, and Alan Borning. 2017. "A Survey of Value Sensitive Design Methods." *Foundations and Trends® in Human-Computer Interaction* 11 (2): 63–125. <https://doi.org/10.1561/11000000015>.
- Friedman, Batya, Peter H Kahn, and Alan Borning. n.d. "Value Sensitive Design: Theory and Methods."

- Friedman, Batya, Peter H. Kahn, Alan Borning, and Alina Hultdtgren. 2013. "Value Sensitive Design and Information Systems." In *Early Engagement and New Technologies: Opening up the Laboratory*, edited by Neelke Doorn, Daan Schuurbiers, Ibo van de Poel, and Michael E. Gorman, 55–95. Philosophy of Engineering and Technology. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-7844-3_4.
- Furlong, Kayla R., Laura N. Anderson, Huiying Kang, Gerald Lebovic, Patricia C. Parkin, Jonathon L. Maguire, Deborah L. O'Connor, Catherine S. Birken, and on behalf of the TARGet Kids! Collaboration. 2016. "BMI-for-Age and Weight-for-Length in Children 0 to 2 Years." *Pediatrics* 138 (1): e20153809. <https://doi.org/10.1542/peds.2015-3809>.
- Genus, Audley, and Andy Stirling. 2018. "Collingridge and the Dilemma of Control: Towards Responsible and Accountable Innovation." *Research Policy* 47 (1): 61–69. <https://doi.org/10.1016/j.respol.2017.09.012>.
- Gibney, Mark, and Peter Haschke. 2020. "Special Issue on Quantitative Human Rights Measures." *Journal of Human Rights* 19 (1): 1–2.
- GiveWell. n.d. "About GiveWell." <https://www.givewell.org/about>.
- Givewell. n.d. "Cost-Effectiveness." <https://www.givewell.org/how-we-work/our-criteria/cost-effectiveness>.
- Glanz, Karen, James F. Sallis, Brian E. Saelens, and Lawrence D. Frank. 2005. "Healthy Nutrition Environments: Concepts and Measures." *American Journal of Health Promotion* 19 (5): 330–33. <https://doi.org/10.4278/0890-1171-19.5.330>.
- Goodhart, Charles. 1975. "Problems of Monetary Management : The U.K. Experience." *Papers in Monetary Economics* 1975 ; 1, Papers in monetary economics 1975 ; 1 ; 1. - [Sydney]. - 1975, p. 1-20, 1.
- Google. n.d. "Google Cloud Model Cards." Accessed June 14, 2023. <https://modelcards.withgoogle.com/about>.
- Greene, Joshua D. 2009. "Dual-Process Morality and the Personal/Impersonal Distinction: A Reply to McGuire, Langdon, Coltheart, and Mackenzie." *Journal of Experimental Social Psychology* 45 (3): 581–84. <https://doi.org/10.1016/j.jesp.2009.01.003>.
- Greene, Joshua D., Leigh E. Nystrom, Andrew D. Engell, John M. Darley, and Jonathan D. Cohen. 2004. "The Neural Bases of Cognitive Conflict and Control in Moral Judgment." *Neuron* 44 (2): 389–400. <https://doi.org/10.1016/j.neuron.2004.09.027>.
- Greene, Joshua D., R. Brian Sommerville, Leigh E. Nystrom, John M. Darley, and Jonathan D. Cohen. 2001. "An FMRI Investigation of Emotional

- Engagement in Moral Judgment.” *Science* 293 (5537): 2105–8. <https://doi.org/10.1126/science.1062872>.
- Gupta, Sanjeev, Estelle Xue Liu, and Carlos Mulas-Granados. 2015. “Politics and Public Investment.” *International Monetary Fund, Finance & Development* 52 (4). <https://www.imf.org/external/pubs/ft/fandd/2015/12/gupta.htm>.
- Hager, Gregory D., Ann Drobnis, Fei Fang, Rayid Ghani, Amy Greenwald, Terah Lyons, David C. Parkes, et al. 2019. “Artificial Intelligence for Social Good.” <https://doi.org/10.48550/ARXIV.1901.05406>.
- Hall, Robert W. 1988. “Plato and Totalitarianism.” *Polis: The Journal for Ancient Greek Political Thought* 7 (2): 105–14. <https://doi.org/10.1163/20512996-90000317>.
- Harnacke, Caroline. 2013. “Disability and Capability: Exploring the Usefulness of Martha Nussbaum’s Capabilities Approach for the UN Disability Rights Convention.” *Journal of Law, Medicine & Ethics* 41 (4): 768–80.
- Helliwell, John F., and Christopher P. Barrington-Leigh. 2010. “Viewpoint: Measuring and Understanding Subjective Well-Being.” *Canadian Journal of Economics/Revue Canadienne d’économique* 43 (3): 729–53. <https://doi.org/10.1111/j.1540-5982.2010.01592.x>.
- Highsmith, John. 2015. “Priorities and Politics.” ThoughtWorks. May 12, 2015. <https://www.thoughtworks.com/en-us/insights/blog/priorities-and-politics>.
- Holland, Isabel, Nicole V. DeVille, Matthew H. E. M. Browning, Ryan M. Buehler, Jaime E. Hart, J. Aaron Hipp, Richard Mitchell, et al. 2021. “Measuring Nature Contact: A Narrative Review.” *International Journal of Environmental Research and Public Health* 18 (8): 4092. <https://doi.org/10.3390/ijerph18084092>.
- “Homepage | RSF.” 2023. June 22, 2023. <https://rsf.org/en>.
- Hsieh, Nien-hê, and Henrik Andersson. 2021. “Incommensurable Values.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2021. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/fall2021/entries/value-incommensurable/>.
- Huffman, Carl. 2019. “Pythagoreanism.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2019. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/fall2019/entries/pythagoreanism/>.
- Innocence Project. n.d. “Key Provisions in Wrongful Conviction Laws.” <https://www.law.umich.edu/special/exoneration/Documents/Key-Provisions-in-Wrongful-Conviction-Compensation-Laws.pdf>.

- James, Aaron. 2005. "Constructing Justice for Existing Practice: Rawls and the Status Quo." *Philosophy & Public Affairs* 33 (3): 281–316.
- Jenkins, Ryan, Kristian Hammond, Sarah Spurlock, and Leilani Gilpin. 2022. "Separating Facts and Evaluation: Motivation, Account, and Learnings from a Novel Approach to Evaluating the Human Impacts of Machine Learning." *AI & SOCIETY*, March. <https://doi.org/10.1007/s00146-022-01417-y>.
- Joseph Saveri Law Firm. n.d. "Facebook Content Moderators' Safe Workplace Litigation." <https://www.saverilawfirm.com/our-cases/facebook-content-moderators-safe-workplace-litigation/>.
- Kahneman, Daniel, and Alan B. Krueger. 2006. "Developments in the Measurement of Subjective Well-Being." *Journal of Economic Perspectives* 20 (1): 3–24. <https://doi.org/10.1257/089533006776526030>.
- Kawamichi, Hiroaki, Sho K. Sugawara, Yuki H. Hamano, Kai Makita, Takanori Kochiyama, and Norihiro Sadato. 2016. "Increased Frequency of Social Interaction Is Associated with Enjoyment Enhancement and Reward System Activation." *Scientific Reports* 6 (1): 24561. <https://doi.org/10.1038/srep24561>.
- Kraft, Jonas. 2017. "The Politics of Investment: How Policy Structure Shapes Political Priorities." PhD Thesis, Aarhus University. https://politica.dk/fil-admin/politica/Dokumenter/Afhandlinger/jonas_kraft.pdf.
- Kregg, Christing. 2016. *Right to Counsel: Mental Health Approaches to Support the Exonerated*. The Univ. of Chicago: Crown Family Sch. Of Social Work, Pol'y, and Practice. <https://crownschool.uchicago.edu/right-counsel-mental-health-approaches-support-exonerated>.
- Lechterman, Theodore M. 2020. "The Effective Altruist's Political Problem." *Polity* 52 (1): 88–115. <https://doi.org/10.1086/706867>.
- Legner, Lauren, and Josh Arndt. 2022. "The Psychological Consequences of a Wrongful Conviction and How Compensation Statutes Can Mitigate the Harms." *Michigan State Law Review*, April. <https://www.michiganstatelaw-review.org/vol-2021-2022/2022/4/25/the-psychological-consequences-of-a-wrongful-conviction-and-how-compensation-statutes-can-mitigate-the-harms>.
- Liu, Lydia T., Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2019. "Delayed Impact of Fair Machine Learning." In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 6196–6200. Macao, China: International Joint Conferences on Artificial Intelligence Organization. <https://doi.org/10.24963/ijcai.2019/862>.
- Lucas, Richard E. 2018. "Reevaluating the Strengths and Weaknesses of Self-Report Measures of Subjective Well-Being." In *Handbook of Well-Being*, edited by Ed Diener, Shigehiro Oishi, and Louis Tay.

- MacIntyre, Alasdair. 1988. *Whose Justice? Which Rationality?* Notre Dame, Ind: Univ. of Notre Dame Press.
- MacIntyre, Alasdair C. 2007. *After Virtue: A Study in Moral Theory*. 3rd ed. Notre Dame, Ind: University of Notre Dame Press.
- Matthews, Joseph. 2021. *How to Win Your Personal Injury Claim*. NOLO.
- Mehrabi, Ninareh, Pei Zhou, Fred Morstatter, Jay Pujara, Xiang Ren, and Aram Galstyan. 2021. "Lawyers Are Dishonest? Quantifying Representational Harms in Commonsense Knowledge Resources." arXiv. <https://doi.org/10.48550/arXiv.2103.11320>.
- message, Send. 2023. "HDCA: Welcome to the HDCA." *HDCA* (blog). June 2023. <https://hd-ca.org/>.
- Murena, Fabio. 2004. "Measuring Air Quality over Large Urban Areas: Development and Application of an Air Pollution Index at the Urban Area of Naples." *Atmospheric Environment* 38 (36): 6195–6202. <https://doi.org/10.1016/j.atmosenv.2004.07.023>.
- Narayan, Deepa, and Patti Petesch. 2002. *Voices of the Poor : From Many Lands*. Washington, DC: World Bank and Oxford University Press. <https://doi.org/10.1596/0-8213-5049-8>.
- Niëns, L.M., and W.B.F. Brouwer. 2013. "Measuring the Affordability of Medicines: Importance and Challenges." *Health Policy* 112 (1–2): 45–52. <https://doi.org/10.1016/j.healthpol.2013.05.018>.
- Nieva, Richard. 2018. "Here's How Facebook Uses Artificial Intelligence to Take down Abusive Posts." CNET. May 2, 2018. <https://www.cnet.com/tech/tech-industry/heres-how-facebook-uses-artificial-intelligence-to-take-down-abusive-posts-f8/>.
- Noble, Safiya. 2018. "Google Has a Striking History of Bias Against Black Girls." Time. March 26, 2018. <https://time.com/5209144/google-search-engine-algorithm-bias-racism/>.
- Nordhaus, William. 1975. *The Political Business Cycle*. Vol. 42. 2.
- Nussbaum, Martha. 2000. "Non-Relative Virtues: An Aristotelian Approach." In *Moral Disagreements*, edited by Christopher Gowans, 1st ed., 278. London: Routledge. <https://www.taylorfrancis.com/books/edit/10.4324/9780203134436/moral-disagreements-christopher-gowans?refId=e935226f-3272-4444-bcd3-114e5476a8c5&context=ubx>.
- Nussbaum, Martha C. 2008. "Creating Capabilities: The Human Development Approach and Its Implementation." *Hypatia* 24 (3): 211–15. <https://doi.org/10.1111/j.1527-2001.2009.01053.x>.

- OKC Injury Lawyer. 2022. "What Are Non-Economic Damages and How Are They Measured?" May 5, 2022. <https://okcinjurylawyer.com/2022/05/what-are-non-economic-damages-and-how-are-they-measured/>.
- O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. First edition. New York: Crown.
- Pavot, William. 2018. "The Cornerstone of Research on Subjective Well-Being: Valid Assessment Methodology." In *Handbook of Well-Being*, edited by Ed Diener, Shigehiro Oishi, and Louis Tay.
- Perrigo, Billy. 2019. "Facebook Says It's Removing More Hate Speech Than Ever Before. But There's a Catch." Time. November 26, 2019. <https://time.com/5739688/facebook-hate-speech-languages/>.
- Plato. 2004. *Republic*. Translated by C. D. C. Reeve. Indianapolis: Hackett Pub. Co.
- Poel, Ibo van de. 2015. "Value Conflict in Design for Values." In *Handbook of Ethics, Values, and Technological Design*, edited by Jeroen van den Hoven, Pieter E. Vermaas, and Ibo van de Poel, 89–116. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-6970-0_5.
- Prieto, Luis, and José A Sacristán. 2003. "Problems and Solutions in Calculating Quality-Adjusted Life Years (QALYs)." *Health and Quality of Life Outcomes*, 8.
- Ranis, Gustav, Frances Stewart, and Emma Samman. 2006. "Human Development: Beyond the Human Development Index." *Journal of Human Development* 7 (3): 323–58. <https://doi.org/10.1080/14649880600815917>.
- Ravetz, Jerome R. 1971. *Scientific Knowledge and Its Social Problems*. Transaction Publishers.
- Raz, Joseph. 2008. "The Claims of Reflective Equilibrium." *Inquiry*, August. <https://doi.org/10.1080/00201748208601970>.
- "Renter Cost Burdens Reach Record Levels | Joint Center for Housing Studies." n.d. Accessed June 24, 2023. <https://www.jchs.harvard.edu/son-2023-cost-burdens-map>.
- Robeyns, Ingrid, and Morten Fibieger Byskov. 2023. "The Capability Approach." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Summer 2023. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/sum2023/entries/capability-approach/>.
- Ross, W. D. 2002. *The Right and the Good*. Edited by Philip Stratton-Lake. New ed. Oxford: Clarendon Press.
- Rousseau, Ronald, and Piet Van Hecke. 1999. "Measuring Biodiversity." *Acta Biotheoretica* 47 (1): 1–5. <https://doi.org/10.1023/A:1002093825480>.

- Rumsfeld, Donald. 2002. "Defense.Gov Transcript: DoD News Briefing - Secretary Rumsfeld and Gen. Myers." U.S. Department of Defense. February 12, 2002. <https://archive.ph/20180320091111/http://archive.defense.gov/Transcripts/Transcript.aspx?TranscriptID=2636>.
- Schokkaert, Erik, and Luc Van Ootegem. 1990. "Sen's Concept of the Living Standard Applied to the Belgian Unemployed." *Recherches Économiques de Louvain* 56 (3-4): 429-50. <https://doi.org/10.1017/S0770451800043980>.
- "Scola v. Facebook FAQs." n.d. Scola v. Facebook Settlement. <https://content-moderatorsettlement.com/Home/FAQ#faq9>.
- Selbst, Andrew D. 2021. "An Institutional View Of Algorithmic Impact Assessments." SSRN Scholarly Paper 3867634. Rochester, NY: Social Science Research Network. <https://papers.ssrn.com/abstract=3867634>.
- Sen, Amartya. 2008. *Commodities and Capabilities*. 13th impr. Oxford India Paperbacks. New Delhi: Oxford Univ. Press.
- Simonite, Tom. 2019. "Algorithms Should've Made Courts More Fair. What Went Wrong?" *Wired*, September. <https://www.wired.com/story/algorithms-shouldve-made-courts-more-fair-what-went-wrong/>.
- Skalse, Joar, Nikolaus H. R. Howe, Dmitrii Krashennnikov, and David Krueger. 2022. "Defining and Characterizing Reward Hacking." <https://doi.org/10.48550/ARXIV.2209.13085>.
- Staghøj, Søren, and Suthan Krishnarajan. 2021. "Who Cares About Free Speech?," *The Future of Free Speech*.
- Sterling, Justin. 2020. "What Methods Are Used to Calculate Pain and Suffering?" *The Sterling Firm*. January 15, 2020. <https://thesterlingfirm.com/methods-used-for-pain-and-suffering/>.
- Strathern, Marilyn. 1997. "'Improving Ratings': Audit in the British University System." *European Review* 5 (3): 305-21. [https://doi.org/10.1002/\(SICI\)1234-981X\(199707\)5:3<305::AID-EURO184>3.0.CO;2-4](https://doi.org/10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4).
- Tabassi, Elham (Fed). n.d. "AI Risk Management Framework: Initial Draft - March 17, 2022," 23.
- Twenge, Jean M., Jonathan Haidt, Andrew B. Blake, Cooper McAllister, Hannah Lemon, and Astrid Le Roy. 2021. "Worldwide Increases in Adolescent Loneliness." *Journal of Adolescence* 93 (1): 257-69. <https://doi.org/10.1016/j.adolescence.2021.06.006>.
- Ueda, Peter, Catherine H. Mercer, Cyrus Ghaznavi, and Debby Herbenick. 2020. "Trends in Frequency of Sexual Activity and Number of Sexual Partners Among Adults Aged 18 to 44 Years in the US, 2000-2018." *JAMA Network Open* 3 (6): e203833. <https://doi.org/10.1001/jamanetworkopen.2020.3833>.

- Umbrello, Steven, and Angelo Frank De Bellis. 2018. "A Value-Sensitive Design Approach to Intelligent Agents." SSRN Scholarly Paper. Rochester, NY. <https://papers.ssrn.com/abstract=3105597>.
- Umbrello, Steven, and Ibo van de Poel. 2021. "Mapping Value Sensitive Design onto AI for Social Good Principles." *AI and Ethics* 1 (3): 283–96. <https://doi.org/10.1007/s43681-021-00038-3>.
- UNDP, ed. 1995. *Human Development Report 1995*. Human Development Report / Publ. for the United Nations Development Programme (UNDP) 1995. New York: Oxford Univ. Press.
- "U.S. Bureau of Labor Statistics." n.d. Accessed June 24, 2023. <https://www.bls.gov/>.
- Valentín-Sívico, Javier. 2022. "Evaluating Barriers To and Impacts Of Rural Broadband Access." Missouri University of Science and Technology.
- Walzer, Michael. 2010. *Spheres of Justice: A Defense of Pluralism and Equality*. Nachdr. New York: Basic Books.
- Wang, Angelina, Solon Barocas, Kristen Laird, and Hanna Wallach. 2022. "Measuring Representational Harms in Image Captioning." arXiv. <http://arxiv.org/abs/2206.07173>.
- Weaver, David, ed. 1998. *The Global Journalist: News People Around the World*. New Jersey: Hampton Press. <https://iamcr.org/publications/hampton/hampton-weaver-1998>.
- Weissbourd, Richard, Milena Batanova, Virginia Lovison, and Eric Torres. n.d. "How the Pandemic Has Deepened an Epidemic of Loneliness and What We Can Do About It." Available at <https://mcc.gse.harvard.edu/reports/loneliness-in-america>. Accessed July 7, 2023.
- Wendt, Claus. 2009. "Mapping European Healthcare Systems: A Comparative Analysis of Financing, Service Provision and Access to Healthcare." *Journal of European Social Policy* 19 (5): 432–45. <https://doi.org/10.1177/0958928709344247>.
- Whitehead, S. J., and S. Ali. 2010. "Health Outcomes in Economic Evaluation: The qaly and Utilities." *British Medical Bulletin* 96 (1): 5–21. <https://doi.org/10.1093/bmb/ldq033>.
- Wiessner, Daniel. 2021. "Judge OKs \$85 Mln Settlement of Facebook Moderators' PTSD Claims." Reuters. July 23, 2021. <https://www.reuters.com/legal/transactional/judge-oks-85-mln-settlement-facebook-moderators-ptsd-claims-2021-07-23/>.
- Wilensky, Harold L. 1964. "The Professionalization of Everyone?" *American Journal of Sociology* 70 (2): 137–58.

- Yale Center for Environmental Law and Policy (YCELP) at Yale University, and Center for International Earth Science Information Network (CIESIN) at Columbia University. 2022. “2022 Environmental Performance Index (EPI).” NASA Socioeconomic Data and Applications Center (SEDAC). 2022. <https://doi.org/10.7927/dwt2-9k25>.
- Yampolskiy, Roman V. 2019. “Predicting Future ai Failures from Historic Examples.” *Foresight* 21 (1): 138–52. <https://doi.org/10.1108/fs-04-2018-0034>.