Toward the AI Index

Yoav Shoham

■ *The AI Index is a new effort to track key developments in AI in a factual and objective way, and in doing so inform discussion and decision making both within AI and outside it. Since the project is early on, the goal of this article is not to present a final product, but rather to convey the current state of the index and invite the community's participation in helping to shape it.*

AI is all the rage these days, and with the attention comes the responsibility — our responsibility, as AI researchers and practitioners — to communicate clearly about our field. What areas are progressing, and at what pace? What are the successes and the challenges? Will AI improve our lives? For example, will it increase productivity, open up new business opportunities, lengthen life expectancy and improve its quality? Or is AI a threat to society? For example, will it eliminate jobs, accentuate biases, or lead to unfair concentration of knowledge and wealth?

Many people are chiming in on these questions. Opinions diverge, which is to be expected in a fast-moving area with more unknowns than knowns. It is an important conversation to have. And precisely because of the inherent uncertainty, it's important that the conversation be anchored in fact. It's hard enough for AI practitioners to keep track of everything that's going on in our exploding field and make

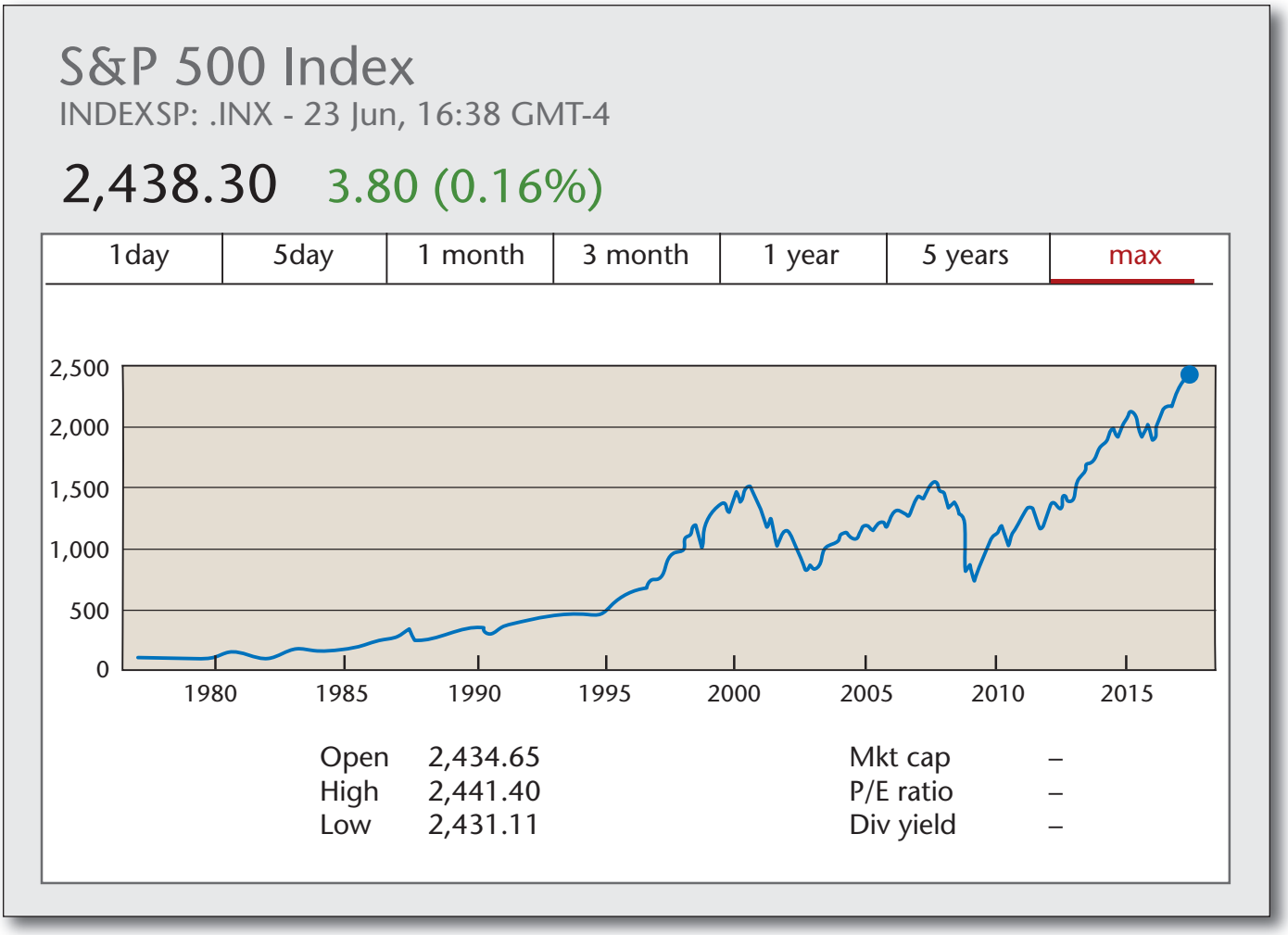

Figure 1. The S&P500 Index.

(Reproduced with permission from S&P Dow Jones Indexes)

sense of it, and it's all the more daunting to outsiders. The goal of the AI Index[1] is to provide precisely this factual basis to the conversation, in an open, not-for-profit fashion.

There are five audiences for which the index is intended: (1) AI scientists and practitioners who are immersed in particular areas but want the big picture; (2) industry leaders who need to decide on strategy and make investment decisions; (3) government that needs to set policy and make funding decisions (the latter is relevant also to private funding bodies); (4) economists and other social scientists who want their analyses of AI's implications to be well-grounded; and (5) the general public, which stands to be the most affected by AI (and the media that serve this public).

The project, which is only a few months old, was conceived under the umbrella of the One Hundred Year Study on Artificial Intelligence (ai100.stanford.edu), nicknamed AI100. The idea was to complement the recurring, in-depth studies commissioned by AI100 with an ongoing window into the field, provide a snapshot at any point in time, and track historical trends. (It's important to mention that while the AI Index was conceived and is being incubated under the AI100 project, its long-term home has not yet been determined.)

Writing this article presents a challenge for timing reasons. There are two phases to the project. The first phase is to define what the index will consist of, how and how often it will be published, and how the project will operate on an ongoing basis (formal home, staffing, governance, funding). The second phase of the project is to operate within the structure defined in the first phase. At the time of writing, the team (more on which at the end) is in the midst of the first phase. By the time the article is published, the project will likely have transitioned to the second phase. So one option would have been to wait until that stage and only write the article then. But since the primary goal of the article is invite the community's involvement in helping evolve and improve the

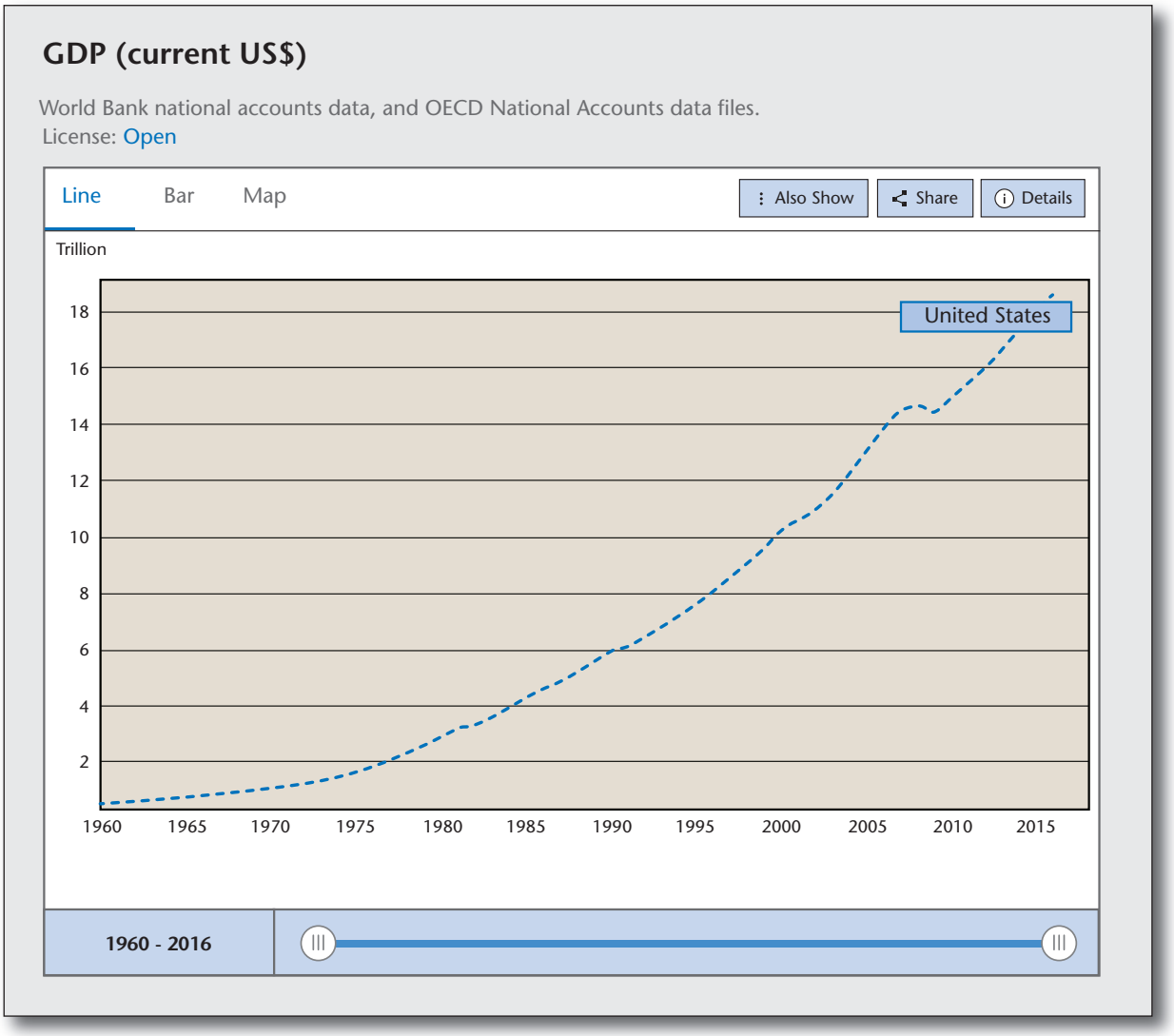

Figure 2. The US Gross Domestic Product.

index, that would be quite later, given publication time lines. So please keep this in mind as your read this article; it will by no means present a finished product, but rather will describe the goals and the state of the effort circa August 2017. Please go to the AI Index home for the most up-to-date status of the index.

About Indexes in General

When an index is mentioned we usually immediately think of a financial index such as the Standard and Poor's 500 financial index, S&P500, created by S&P Dow Jones Indexes (figure 1). We also think of an economic measures index, such as the gross domestic product (GDP), compiled by the US Department of Commerce, Bureau of Economic Analysis (figure 2).

However, there are scores of other indexes, which vary both in the subject matter and the form of the index. The S&P and the GDP are both economic measures, and consist of a single number tracked over time. But take, for example, the Social Progress Index, created by the Social Progress Imperative and made up of 54 different components that capture three dimensions of social progress: basic human needs, foundations of wellbeing, and opportunity (figure 3). These are then made into a composite index, and countries are compared qualitatively in a visual map (figure 4).

As a final and very different example, there is Middle East Peace Index, which consists of monthly surveys that change over time so as to best capture current affairs. The poll results are communicated along with analysis and commentary, but no attempt is made to make a composite measure out of the results.

So the design space of indexes is very broad. An index can be objective or subjective; consist of a single score or many; and may contain a textual component that is an indispensible part of the index.

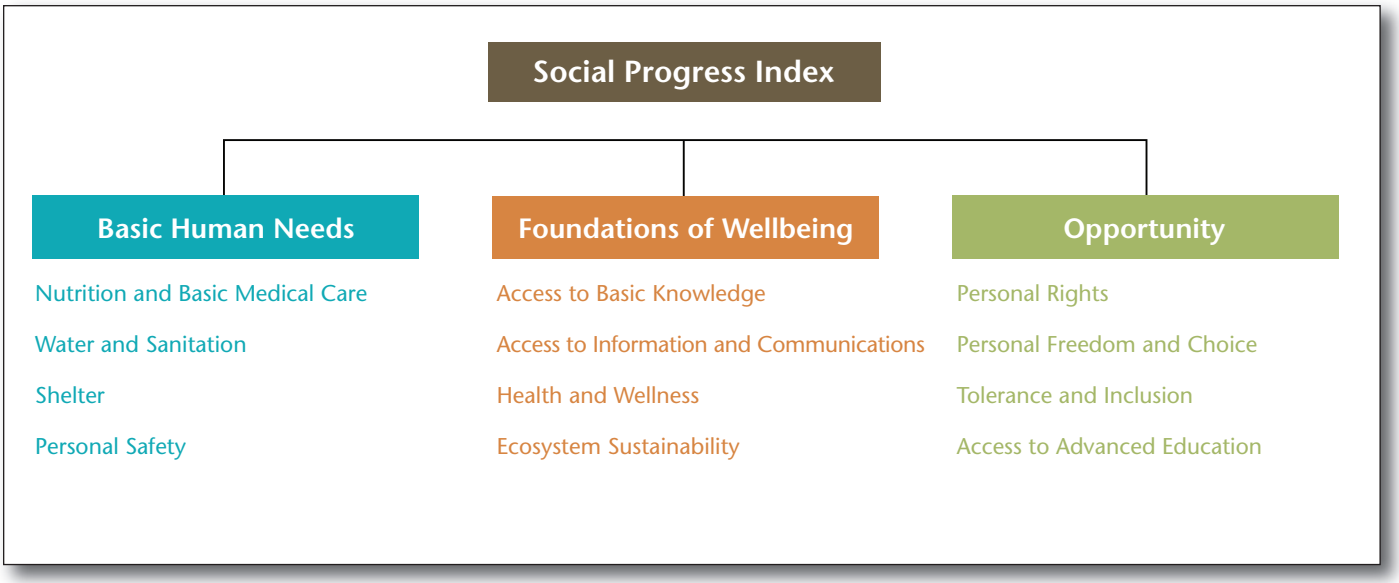

Figure 3. Some Dimensions of the Social Progress Index.

(Reproduced with permission from the Social Progress Imperative)

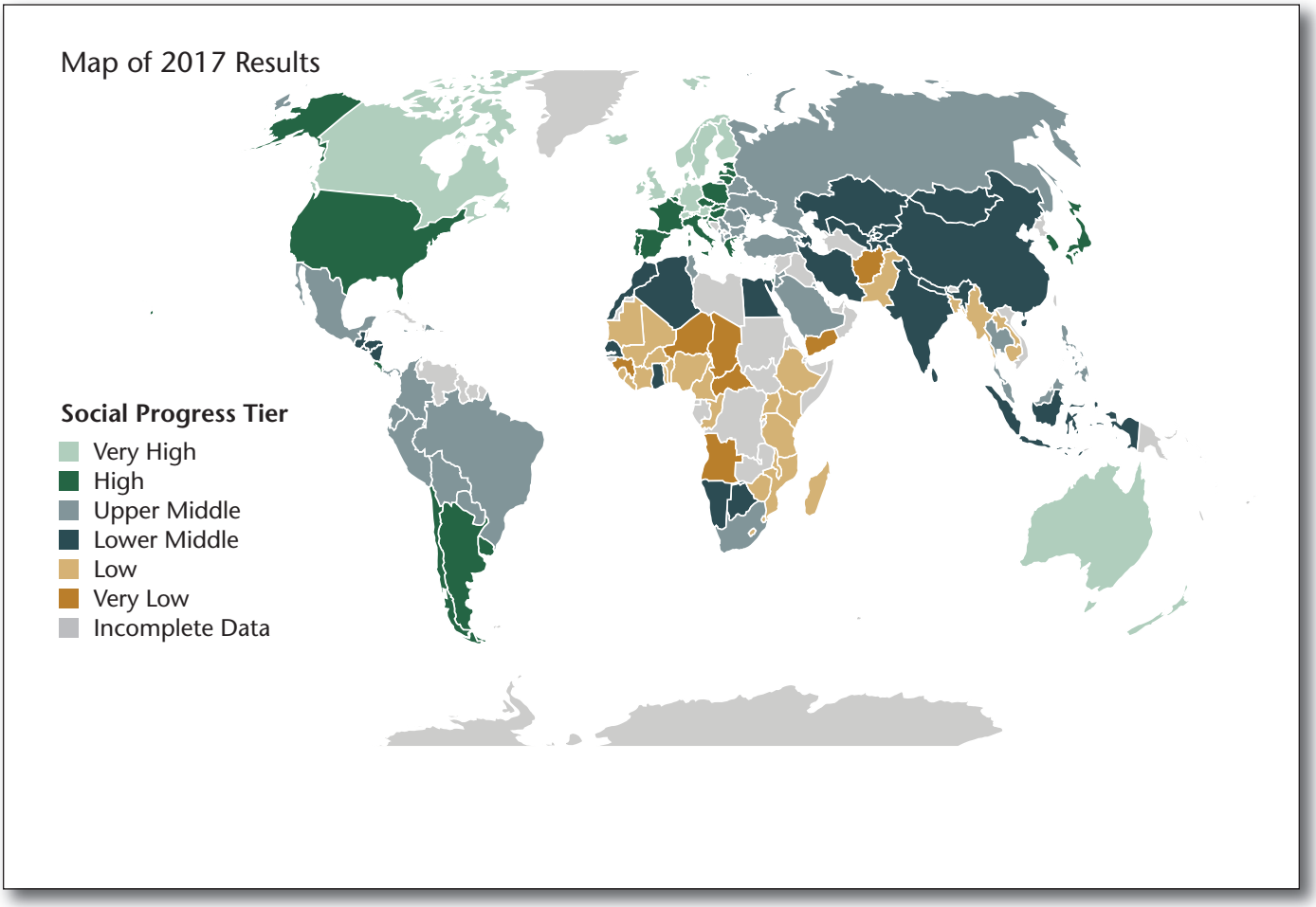

Figure 4. The Consolidated Social Progress Index, Map View.

(Reproduced with permission from the Social Progress Imperative)

However, for all their diversity, all indexes have the same goal. They attempt to take a domain (economy, peace, AI, and so on) that is complex and hard to grasp — certainly by lay people, but also by the experts — and extract from it a simple core that captures the essence. By definition, an index loses information about the domain; in fact, it loses most of the information. And so, if constructed poorly, an index can be very misleading. Indeed, even the most venerable measures such as the GDP are often criticized for this reason. But well-constructed indexes are indispensible; without them it's just impossible for all but a few to understand the domain. The index helps frame the conversation, provides a meaningful if not full or even fully accurate picture of the domain, and tracks changes over time. It helps practitioners see the forest for the trees, policy makers decide on policy, business executives to plan strategy, and not the least the general public understand a domain that affects it profoundly.

So how does one design an index that is useful rather than misleading? From my experience this happens in three stages. First, domain experts take an initial, informed stab at it (1). Next, statistical methodology is applied to avoid gross errors with regard to aspects such as data selection bias and ways of aggregating scores from different scales into a composite score (2). Third, the index is honed over time based on experience and feedback by the community (3), and then one of two things happens: Either the index takes hold and is used broadly (3a), or it withers on the vine because it's viewed as either misleading or irrelevant or both (3b).

The AI Index is following the same path, with the hope of ending up in being used broadly (outcome 3a rather than(3b), shepharded by an organization that continuously maintains and improves the index.

What Will the AI Index Track?

As I emphasized at the beginning, the project is early on, so it's premature to describe the content of the AI Index with any definiteness. That said, there are some general things that can be said about the content. We see three primary dimensions that the index should in principle cover: (1) volume of activity; (2) technological progress; and (3) societal impact.

Of these, the first dimension is the simplest to both define and measure. Its goal is to capture how vibrant the area is, perhaps with some indication of relative "hot" subfields. So for example we envision tracking the number of attendees and papers at conferences, the amount of venture capital invested in AI startups, and number of AI job postings.

Note that already here we encounter a challenge of what to include. Our rule of thumb is to aim for representativeness rather than comprehensiveness. Take conference attendance for example. We will start with a set of conferences that are widely viewed as core to AI and leaders in their categories; this will likely include the AAAI conference of the Association for the Advancement of Artificial Intelligence; the International Joint Conferences of Artificial Intelligence (IJCAI); the Conference on Uncertainty in Artificial Intelligence (UAI); the Conference on Neural Information Processing Systems (NIPS); the International Conference on Machine Learning (ICML), the International Conference on Principles and Practice of Constraint Programming (CP); the International Conference on Automated Planning and Scheduling (ICAPS), the International Conference on Autonomous Agents and Multiagent Systems (AAMAS) and the International Conference on Principles of Knowledge Representation and Reasoning (KR) (although even here some controversy is to be expected around conferences in areas that are currently out of favor or perhaps not as central to AI in the eyes of some). Then there are domain-specific conferences where the domains are primary showcases for AI (such as the Annual Meeting of the Association for Computational Linguistics (ACL) for computational linguistics; the International Conference on Computer Vision (ICCV) and the IEEE Conference on Computer Vision and Pattern Recognition for machine vision, or the IEEE International Conference on Intelligent Robots and Systems (IROS) and the IEEE International Conference on Robotics and Automation (ICRA) for robotics). Our current sense is that by now the general AI conferences don't faithfully capture trends in these three areas (linguistics, vision, robotics) so we do need to include these domain-specific conferences. In other cases we don't feel this is required, even though the conferences are of high quality (for example, this likely includes the ACM Conference on Knowledge Discovery and Data Mining (KDD), the International Word Wide Web Conference (WWW), and some statistics conferences). So even a seemingly innocuous measure such as conference attendance calls for some subjective decisions, and similar judgment calls will be needed when we look at AI investments and job postings.

To capture the trendiness of different areas we can break down the above by subarea; one can even imagine a word cloud with salient technical terms. So, for example, one can expect that among other things this component will show that in 1985 "knowledge representation" figured prominently, and in 2015 AI "machine learning" became a leading area. Undoubtedly more nuanced insights will emerge as well, as citation analyses and other bibliometric techniques are applied.

The second dimension — technological progress in various areas of AI — will be harder to pin down, and more important. Again I should emphasize that we are in the process of defining the set of measure to track. But with this caveat, to give a concrete sense, here are some leading candidate areas in which we plan to measure progress:

Image Understanding
Video Understanding
Speech and Natural Language Processing
Information Retrieval
Machine Translation
Dialogue Systems
Satisfiability
Planning
Knowledge Representation, including knowledge graph statistics and more

Again, these candidate areas are illustrative and will surely evolve by the time the initial index is defined. The index will also track progress in certain key application domains that combine performance on the more specific technical areas, such as self-driving cars, game playing, or conversational bots. Much will depend on the availability of data and its characteristics. This brings up serious methodological challenges, more on which will follow in the next section.

The third dimension is aimed at tracking the impact of AI on society, be it on employment and other economic aspects, finance, medicine, education, transportation, government, military, and beyond. If the first two dimensions focus on the production side of AI, this third one focuses on the consumption side. Consumption is arguably the most important aspect to understand, and the hardest. It's not clear what to track, the data is highly diffuse, and the problem of "credit or blame assignment" (to what degree AI is responsible for changes taking place) looms large. We felt that at this stage tackling this aspect would cause the project to grind to a halt and decided to forego it in the first versions of the AI Index. At a minimum, delivering on the first two components of the index will provide data for other researchers to do more complex and grounded analyses of the societal implications. For now, the only nod in this direction that we feel comfortable giving at this stage is some gauge of public interest in AI, both the level of interest (for example, as indicated by Google Trends) and perhaps some measure of "sentiment analysis" in the general media.

Finally, we are considering including an element of subjective, expert commentary. We imagine that the Index will be published often (perhaps continually), but that at some cadence (for example, annually) a report will be issued that will present the findings for the period. This could be an opportunity for a panel of experts to provide commentary on it, add information not captured by the Index, and perhaps make predictions on where AI is headed.

Challenges

Designing an AI index is not a trivial intellectual task. We're not aware of a similar effort to track a scientific or technological area. There are some measures of specific aspects of an area, but not of the entire area. For example, Moore's law has been very influential, but of course it hardly captures all aspects of progress in hardware development. We obviously think attempting a broad index is a worthwhile effort, but we're not blind to the challenges. Some of them follow.

A first challenge is the availability of data. There are some well-established benchmarks in certain area such as vision, machine learning, satisfiability, and planning. But there's the "drunk and lamppost" danger of ignoring key areas in which such benchmarks are lacking. Here again I remind that our approach is to aim for representativeness rather than exhaustiveness. We hope to end with a set of pillars that together span AI reasonably well. And if we (as a community — more on this below) feel that certain key areas aren't represented, to help catalyze an effort to create benchmarks in these areas.

A second challenge is the instability and discontinuity of the data. Financial indexes are an inspiration, but can be misleading if taken literally. In a fast-moving area the benchmarks are a moving target for two reasons, one shallow and one deep.

In certain areas (such as the annual propositional satisfiability problem [SAT] competition) the benchmarks are changed simply because they weren't established with an eye toward tracking progress in a reliable, quantified way. Our philosophy here is twofold. First, where a stable signal can be extracted from the existing data, whether by insights into the domain or algorithmically, to help extract it. And second, when this isn't possible, to work with the community to establish a more stable benchmark.

More fundamentally, new areas emerge that simply didn't exist previously, and existing methods and data sets inevitably become obsolete as knowledge and technology advance. Here we will need to continually revisit the components of the index, and update them to reflect the current status of the field. In a sense this isn't that radical; the S&P 500 routinely swaps stocks in and out of the index. But this will be trickier in the case of the AI Index. First, it's likely that more subjective judgment calls will be needed than in the case of the S&P. But more deeply we will need to embrace discontinuity. We imagine a framework of "punctuated continuity" whereby for a period of time progress is tracked in a uniform, measureable way, and at some point the measure is replaced by a new one, since the old one has been "solved away." Rather than be dismayed by this, we should note and celebrate this achievement, and start tracking the new measures.

Finally, there's a challenge of creating a composite index out of a heterogeneous set of data. To take a (relatively) simple example, how do you turn the number of conference attendees, venture capital investment dollars, and the number of job openings into a composite measure of level of activity? Or, how do you roll up progress in different facets of machine learning into a composite measure of machine-learning progress? There is a certain methodology of index

composition that provides guidance, but the guidance is partial, and care will be needed in how we apply it.

A Community Effort

The AI Index will succeed only if it becomes a community effort. The current team is taking an initial stab, but the index must reflect the general wisdom and engagement of the community. Given the current interest in AI, many reports on AI appear, such as the popular Import AI newsletter.[2] There are also specific efforts to collate AI performance metrics, such as the recent project at the Electronic Frontier Foundation.[3] Although none of these efforts are aimed at curating a subset of the data that is representative of the entire field, let alone building an index, the raw data collected there provides essential input to the AI Index. We aim for a common data repository, open collaboration, and avoiding duplication of effort.

It's not only these specific efforts. We invite all AI scientists and practitioners and policymakers to contribute to the effort. While we imagine our relationship with the broader community will evolve over time, at this stage the AI Index website[4] provides information about some ways community members can get involved. We welcome community members to share data, recommend data sources to track, share domain expertise, and provide general comments and advice to the organizing team. We also encourage all interested parties to sign up to receive updates about the Index on our website and invite the community to send any further questions and comments to community@AIindex.org.

Origins and Startup Phase of the AI Index

I had the privilege of being a founding member of the One Hundred Year Study on AI standing committee, from which I "termed out" earlier this year. The AI Index was conceived during that time, being in resonance with the mission of AI100. I thank the other members of the inaugural standing committee — Barbara Grosz, Eric Horvitz, Alan Mackworth, Tom Mitchell, and Deirdre Mulligan — for their enthusiastic support and wise counsel.

The initial phase of the AI Index project is run out of Stanford University, where Professor Russ Altman, the faculty director of AI100, serves as the official host (he is also member of the AI100 standing committee, ex officio).[5] Russ Altman too deserves many thanks for his help and ideas. (As mentioned earlier, no decision has been made regarding the ultimate home of the AI Index.)

The initial phase of the AI Index is expected to last through the end of 2017 and is led by a steering committee consisting of Ray Perrault from SRI International, Erik Brynjolfsson from the Massachusetts Institute of Technology, Jack Clark from OpenAI, and me. (Ray Perrault is to be thanked in particular for volunteering to help oversee the process). In addition, Calvin LeGassick recently joined as project manager, and we are also fortunate to have expert advice on index creation by Hagar Tzameret from Sapir College. This core team is advised by a larger advisory committee, whose members currently include Michael Bowling, Ernie Davis, Julia Hirschberg, Eric Horvitz, Karen Levy, Alan Mackworth, Chris Manning, Tom Mitchell, Sandy Pentland, Chris Re, Daniela Rus, Sebastian Thrun, Hal Varian, and Toby Walsh.

The AI Index gratefully acknowledges the early financial support of the AI Index by the AI100 project, Google, Microsoft, and Toutiao.

Acknowledgments

I thank Erik Brynjolfsson, Jack Clark, Barbara Grosz, Eric Horvitz, Calvin LeGassick, and Ray Perrault for helpful comments about this article, although of course I alone am responsible for anything that's terrible about the result.

Notes

1. www.AIindex.org.
2. jack-clark.net/import-ai.
3. www.eff.org/ai/metrics.
4. www.AIindex.org.
5. The remainder of this section is written in present tense, although by the time the article is published much of it will be retrospective.

Yoav Shoham is a professor (emeritus) of computer science at Stanford University and a senior scientist at Google. He is a fellow of AAAI, the ACM, and the Game Theory Society, and a serial entrepreneur. He can be contacted at shoham@cs.stanford.edu.