AI Bookie

At ai.sciencebets.org you can make your own predictions, challenge another prediction and turn it into a bet, or post a bet of your own. We are here to help. So, place your bets at ai.sciencebets.org!

■ *The AI Bookie column documents highlights from AI Bets, an online forum for the creation of adjudicatable predictions, in the form of bets, about the future of AI. While it is easy to make broad, generalized, or off-the-cuff predictions about the future, it is more difficult to develop predictions that are carefully thought out, concrete, and measurable. This forum was created to help researchers craft predictions whose accuracy can be clearly and unambiguously judged when the bets come due. The bets will be documented both online and regularly in this column. We encourage bets that are rigorously and scientifically argued. We discourage bets that are too general to be evaluated or too specific to an individual or institution. The goal is not to continue to feed the media frenzy and outsized pundit predictions about AI, but rather to curate and promote bets whose outcomes will provide useful feedback to the scientific community. For detailed guidelines and to place bets, visit sciencebets.org.*

Place Your Bets: The Future of Voice Interfaces to Intelligent Assistants

Chris Welty, Lora Aroyo, Eric Horvitz

Scientific bets come in several forms, two of which we will explore in this issue's column. The first is a genuine disagreement between scientists that arises from scientific discourse, followed by the exercise to codify their disagreement in a measurable way. The second is a set of specific predictions, which are one-sided bets made by a scientist with the intention of inviting other scientists to challenge and formalize the bet.

We begin with a disagreement between two scientists, Chris Welty and Lora Aroyo, about the future of voice interfaces to intelligent assistants. This discussion is followed by a series of one-sided bets from former AAAI president Eric Horvitz.

The Bet

At issue: The speech-driven reality presented in the movie *Her* offers a realistic view — or not — of the foreseeable future (such as 2025).

For: Lora Aroyo

I bet that the speech-driven reality presented in the movie *Her* is a realistic view of the foreseeable future. Here are some arguments in support of this position.

We already have mature hardware technology that can provide a reasonable user experience with a speech-controlled UI in the form of wireless devices with an integrated assistant (for example, Google Assistant, Siri, Alexa). Also, the relevant software technology evolves quickly in terms of accuracy and coverage of speech processing (for example, background noise and strong accents no longer present substantial obstacles for speech understanding).

Speech is a convenient interaction medium because it doesn't require the user to hold or tap anything, and so speech interfaces allow for parallelization with other activities such as cooking, driving, washing dishes, or painting. So far, convenience has been a major factor in the adoption of initially imperfect systems, and this early use inevitably drives further improvement. As the number of early adopters grows, we more quickly reach the tipping point for large-scale adoption. In the interim, those initial users provide data to demonstrate both that this is a continuously growing market and that there is additional value in investing in the implementation of the technology of this market, in this case various types of interactions with virtual assistants.

Wide-scale usage is also going to drive innovation in ML research in terms of finding ways to optimize and generalize the process of creating new interactions, rather than implementing all of them from scratch. Additionally, as more people start using speech interfaces, there will be an increased incentive for researchers to advance new areas of research, confronting barriers in terms of both hardware and software (for example, blocking sound from other people talking in the same space).

Google's search engine broke through the dense search market of the time with its simplicity, that is, by being focused on a single utility: search. It was easy, unambiguous, and convenient for people to use. This simplicity encouraged a steady growth in numbers of users, which meant also a steady increase in the amount of usage data, and all that data, in turn, eventually made the results much, much better.

We have all the paving stones now for the road to simple and convenient speech-driven interaction. And, yes, it is already quite noisy and annoying when people talk on their phones in public spaces. But there is no stopping it, so we may as well adapt. We talk on our phones or into our headsets, while walking and while traveling, on the sidewalks and in the streets, in buses, in trains, and in other public places. We do it because it's convenient: you don't have to take anything out of your bag, you don't have to hold anything in your hands — you just talk. As for adapting, many people are already walking on the street or working at their desk with noise cancellation headsets. This current behavior makes it even easier to adopt speech interaction with assistants in public spaces.

There are some counterexamples. For example, it would be annoying if everybody at home talked to their devices and didn't communicate with one another. Just as with texting and social media we gained the ability to communicate instantly, at any time, but also developed new social norms for when and how to use that ability (or at least some of us have), I believe that in the same way, we will develop new social norms for where and how to use increased voice interaction.

Against: Chris Welty

In the movie *Her*, there is a scene in which the main character is coming home from work and walking across a large outdoor plaza. He is talking to his AI assistant (and lover) through some kind of wireless headset device in his ear. Across the plaza, there are many people who also appear to be leaving work and who are also talking on their headsets, presumably to their computers. My reaction to this scene was that it depicted an improbable future. Initially, my primary reason was simply that I thought the social pressure would prevent this scenario from becoming widespread, just as the social pressure not to talk to someone on your mobile phone in public prevents most of us from doing it and leaves us feeling annoyed when others do.

I expressed my skepticism to Lora Aroyo, who, on the contrary, found this to be a nearly certain future. We started an informal bet at that time, for and against the prediction that *assistant voice interfaces would soon become mainstream* — with "against" meaning that most people would continue instead to use keyboards, mice, and touchscreens to interact with machines. At the instigation of the AI Bookies column, we encouraged ourselves to go through the process of formalizing the bet.

Like many bets and predictions, the need to avoid an open-ended condition drove us to specify a time limit — the year 2025.

The first obstacle we encountered was turning *mainstream* into something measurable. What objective criteria would we use? As I began to think about my motivations and to think about the problem in more concrete provable/disprovable terms, I began to think of more rigorous and serious reasons why I believe speech will never become a mainstream interface.

While speech is a convenient medium for humans to use when interacting with each other, we use it

because we don't have anything better — such as buttons. If there were a button to press to get a kid to clean their room, no parent would waste time with words. Speech is indeed mostly a waste of time.

It's a waste of time because a lot of speech is negotiation. If I want someone to do something, I have to do more than just tell them. I have to negotiate with them using sticks or carrots until either they agree to do it or I move on to find someone else. *Star Trek's* Picard's "make it so" works only in special circumstances and organizational structures. With an assistant, this negotiation is not necessary, but there is a very important second part to the negotiation: the meaning. And therein lies the rub.

Speech is a lousy way to communicate meaning. When designing applications that actually *do* the things we want our assistants to do, a lot of work goes into the user interface. Arguably, it was Google's user interface, not the quality of the search results, that won the day during the search engine wars. Google's interface was *simple* — just a text box. But an assistant is much more than a search box, and each bit of functionality needs to be implemented and the interface designed. Take a simple example: ordering a taxi using Uber or Lyft or the equivalent. I could say to my assistant, "Hey, Google, order me a taxi." "Where are you going?" it would ask. "To Matsui's Sushi." "Do you mean Matsui's hair salon or Matsui's Japanese restaurant?" "The restaurant, obviously, you idiot!" "Great. Your taxi will arrive in five minutes." My obnoxiousness notwithstanding, Matsui's Japanese restaurant is 140 miles away. I don't realize this because I've mistaken the name "Matsui" for "Masuashi," and I've just unintentionally ordered an 140-mile taxi ride. Now maybe, hypothetically, the assistant here is *really* smart, and it realizes something is unusual, and instead of ordering the taxi it says, "Matsui's Japanese restaurant is 140 miles away. Are you sure you want me to order a taxi?" At which point I say, "No!" and perhaps I continue negotiating with my assistant about it, or perhaps it knows my history and enough about speech similarity and American confusions about Japanese names to figure out what I meant. But none of this intelligence would be necessary if I were using the right interface for ordering a taxi: a map! At some point pretty early on — depending, of course, on my mobility and access to my hands and so forth — the negotiation of the meaning becomes so inefficient that I give up and switch to the actual app that provides a well-honed interface to the thing I want done.

Negotiating meaning is certainly an interesting and very hard problem, one studied and emphasized in the past but less so today, and one reason for that decline in emphasis is economic. There exist cheaper alternative solutions that make it unlikely that we will ever have a solution that annoys us so much and that is so difficult to work with.

There are clearly some things for which a speech interface is effective (for example, question answering), and certain conditions for which it is the best possible option (for example, while driving). Furthermore, there has been impressive progress in both speech understanding and speech generation recently. However, speech will not become the primary interface to our assistants and devices simply because there are far more examples of cases for which it is not the best solution, and many for which it is simply terrible.

Adjudicating the Bet

We discussed many aspects of this disagreement with the other bookies, who will act as adjudicators, to help us hone down the adjuticatable portion of the bet. Ultimately, the bettors agreed that some representative of the number of minutes spent talking to devices in a year, normalized by the number of devices available, would be the metric we are looking for. Welty argues that a graph of this usage ratio over the next few years until 2025 will never be more than linear. Aroyo argues that it will reach a tipping point, defined by a superlinear bend in the usage curve before that time.

There are several sources that might be used to provide an approximation of this metric. TechCrunch and SearchEngineLand report on the assistant device market, giving us a normalization factor. Voicebot.ai and alphametic.com report on aspects of the speech understanding industry, and can perhaps be influenced to gather data about the amount of time spent speaking to devices. At the present time, we couldn't find precisely the data we want being gathered today, so we will report back in the next issue.

Eric Horvitz's Predictions

To help stimulate more participation in AI bets, former AAAI president Eric Horvitz contributed the following predictions as one-sided bets. Readers are encouraged to take him on. Challenge one of these predictions and learn something valuable in the process of turning the prediction into something adjuticatable.

Prediction

Laws will require AI systems that emulate humans to reveal themselves.

By 2025, laws will be in place in some parts of the world requiring that systems reveal to people the use of AI in pure or hybrid (human in the loop) AI systems that emulate humans.

Prediction

Artists will certify that their art has been created by humans.

By 2035, a wave of artwork, including painting, poetry, and music, mostly or wholly created by AI systems, will lead some artists to certify that their cre-

Illustration courtesy James Gary.

ations have been created by humans without significant AI assistance.

Prediction

European Union or United States laws will regulate face recognition for surveillance.

By 2025, the use of face recognition for broad surveillance will come under regulatory guidance in the United States or Europe, limiting uses of face recognition by governments for general, large-scale detection of people.

Prediction

Laws will require self-driving cars to share performance and safety data.

By 2030, some countries will require that manufacturers share data from semiautonomous and autonomous vehicles on performance and accidents, akin to the National Transportation Safety Board in the United States for air transport.

Prediction

A self-driving car will cross the continental United States.

By 2025, the first car to cross the United States without human intervention will be celebrated.

There will be roads closed to human drivers. By 2030, at least one city in the world will close a region to all human-piloted vehicles and employ a mix of autonomous cars, including large-scale public micro-transit systems.

Prediction

Governments will start to build road infrastructure for self-driving cars.

By 2035, in numerous regions in the United States, Europe, and elsewhere, specially designed signaling and related infrastructure will be deployed that allows semiautonomous vehicles to be reliably autonomous for long stretches of highway. Special "hyperlanes" will be created in places that allow for high-speed coordinated travel.

Chris Welty is a senior research scientist at Google, Inc.

Lora Aroyo is a computer scientist and professor at The Vrije Universiteit Amsterdam, Netherlands.

Eric Horvitz is a technical fellow and director at Microsoft Research.