Volume Title: The Economics of Artificial Intelligence: An Agenda

Volume Authors/Editors: Ajay Agrawal, Joshua Gans, and Avi Goldfarb, editors

Volume Publisher: University of Chicago Press

Volume ISBNs: 978-0-226-61333-8 (cloth); 978-0-226-61347-5 (electronic)

Volume URL: http://www.nber.org/books/agra-1

Conference Date: September 13–14, 2017

Publication Date: May 2019

Chapter Title: Comment on "Artificial Intelligence and Behavioral Economics"

Chapter Author(s): Daniel Kahneman

Chapter URL: http://www.nber.org/chapters/c14016

Chapter pages in book: (p. 608 – 610)

ics, edited by J. Kagel and A. Roth, 253–348. Princeton, NJ: Princeton University Press.

Rumelhart, D. E., and J. L. McClelland. 1986. "On Learning the Past Tenses of English Verbs." In *Parallel Distributed Processing*, vol. 2, edited by D. Rumelhart, J. McClelland, and the PDP Research Group, 216–71. Cambridge, MA: MIT Press.

Sawyer, J. 1966. "Measurement and Prediction, Clinical and Statistical." *Psychological Bulletin* 66:178–200.

Stahl, D. O., and P. W. Wilson. 1995. "On Players' Models of Other Players: Theory and Experimental Evidence." *Games and Economic Behavior* 10 (1): 218–54.

Thornton, B. 1977. "Linear Prediction of Marital Happiness: A Replication." *Personality and Social Psychology Bulletin* 3:674–76.

Tversky, A., and D. Kahneman. 1992. "Advances in Prospect Theory: Cumulative Representation of Uncertainty." *Journal of Risk and Uncertainty* 5 (4): 297–323.

von Winterfeldt, D., and W. Edwards. 1973. "Flat Maxima in Linear Optimization Models." Working Paper no. 011313-4-T, Engineering Psychology Lab, University of Michigan, Ann Arbor.

Wang, J., M. Spezio, and C. F. Camerer. 2010. "Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth Telling and Deception in Sender-Receiver Games." *American Economic Review* 100 (3): 984–1007.

Wright, J. R., and K. Leyton-Brown. 2014. "Level-0 Meta-models for Predicting Human Behavior in Games." In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, 857–74.

Comment Daniel Kahneman

Below is a slightly edited version of Professor Kahneman's spoken remarks.

During the talks yesterday, I couldn't understand most of what was going on, and yet I had the feeling that I was learning a lot. I will have some remarks about Colin (Camerer) and then some remarks about the few things that I noticed yesterday that I could understand.

Colin had a lovely idea that I agree with. It is that if you have a mass of data and you use deep learning, you will find out much more than your theory is designed to explain. And I would hope that machine learning can be a source of hypotheses. That is, that some of these variables that you identify are genuinely interesting.

At least in my field, the bar for successful publishable science is very low. We consider theories confirmed even when they explain very little of the variance so long as they yield statistically significant predictions. We treat the residual variance as noise, so a deeper look into the residual variance, which machine learning is good at, is an advantage. So as an outsider, actu-

Daniel Kahneman is professor emeritus of psychology and public affairs at the Woodrow Wilson School and the Eugene Higgins Professor of Psychology emeritus, Princeton University, and a fellow of the Center for Rationality at the Hebrew University in Jerusalem.

For acknowledgments, sources of research support, and disclosure of the author's material financial relationships, if any, please see http://www.nber.org/chapters/c14016.ack.

ally, I was surprised not to hear more about that aspect of the superiority of artificial intelligence (AI) compared to what people can do. Perhaps, as a psychologist, this is what interests me most. I'm not sure that new signals will always be interesting, but I suppose that some may lead to new theory and that would be useful.

I do not fully agree with Colin's second idea: that it is useful to view human intelligence as a weak version of artificial intelligence. There certainly are similarities, and certainly you can model some of human overconfidence in that way. But I do think that the processes that occur in human judgment are quite different than the processes that produce overconfidence in software.

Now I turn to some general remarks of my own based on what I learned yesterday. One of the recurrent issues, both in talks and in conversations, was whether AI could eventually do whatever people can do. Will there be anything that is reserved for human beings?

Frankly, I don't see any reason to set limits on what AI can do. We have in our heads a wonderful computer. It is made of meat, but it's a computer. It's extremely noisy, but it does parallel processing. It is extraordinarily efficient, but there is no magic there. So, it is difficult to imagine that, with sufficient data in the future, there will remain things that only humans can do.

The reason that we see so many limitations, I think, is that this field is really at the very beginning. I mean we are talking about developments (i.e., deep learning) that took off eight years ago. That is nothing. You have to imagine what it might be like in fifty years. Because the one thing that I find extraordinarily surprising in what is happening in AI these days is that everything is happening faster than we expected. People were saying that it will take ten years for AI to beat Go. The interesting thing is it took less by an order of magnitude. This excess of speed at which this thing is developing and accelerating, I think, is very remarkable. So, setting limits is certainly premature.

One point that was made yesterday was about the uniqueness of humans when it comes to evaluations. It was called judgment, but in my jargon it is "evaluation." Evaluations of outcomes are, basically, the utility side of the decision function. I do not see why that should be reserved for humans. On the contrary, I would like to make the following argument: the main characteristic of people is that they are very noisy. You show them the same stimulus twice and they do not give you the same response twice. We have stochastic choice theory because there is so much variability in people's choices conditional on the same stimuli. What can be done with AI is to create a program that observes an individual's choices. That program will be better than people at a wide variety of things. In particular, it will make better choices for the individual. Why? Because it will be noise free. We know from the literature that Colin cited on predictions that there is an interesting tidbit. Take some clinicians and have them predict some criterion a large number of times. Then develop a simple equation that predicts, not the outcome, but each clinician's judgment. That model does better in predicting the outcome than the clinicians themselves.

That is fundamental. It is telling you that one of the major limitations on human performance is not bias, it is just noise. I may be partly responsible for this as, when people now talk about error, they tend to think of bias as an explanation. That's the first thing that comes to mind when there is an error in human performance.

In fact, most of the errors that people make are better viewed as random noise, and there is an awful lot of it. Admitting the existence of noise has implications for practice. One implication is obvious. You should replace humans by algorithms whenever possible. Even when the algorithm does not do very well, humans do so poorly and are so noisy that, just by removing the noise, you can do better than people. The other is that when you cannot replace the human by an algorithm, you try to have human simulate an algorithm. The idea is that, by enforcing regularity, process and discipline on judgment and on choice, you reduce the noise, and you improve performance because noise is so pernicious.

Yann LeCun said yesterday that humans would always prefer emotional contact with other humans. That strikes me as probably wrong. It is extremely easy to develop stimuli to which people will respond emotionally. An expressive face that changes expressions, especially if it's baby-shaped, gives cues that will make people feel very emotional. Robots will have these cues. Furthermore, it is already the case that AI reads faces better than people do. Undoubtedly, robots will be able to predict emotions and development in emotions far better than people can.

I really can imagine that one of the major uses of robots will be taking care of the old. I can imagine that many old people will prefer to be taken care of by friendly robots that have a name, have a personality, and are always pleasant. They will prefer that to being taken care of by their children.

I want to end on a story. A well-known novelist wrote me some time ago that he's planning a novel. The novel is about a love triangle between two humans and a robot. What he wanted to know is how the robot would be different from the people.

I proposed three main differences. One is obvious: the robot will be much better at statistical reasoning and less enamored with stories and narratives than people are. The other is that the robot would have a much higher emotional intelligence. The third is that the robot would be wiser. Wisdom is breadth. Wisdom is not having too narrow a view. That is the essence of wisdom; it's broad framing. A robot will be endowed with broad framing. I say that when it has learned enough, it will be wiser than we people because we do not have broad faming. We are narrow thinkers, we are noisy thinkers, and it is very easy to improve upon us. I do not think that there is very much that we can do that computer will not eventually be programmed to do.