AI as Influencer

To the Editor:

The paper on "Trusting Intelligent Machines: Deepening Trust Within Socio-Technical Systems" (*IEEE Technology and Society Magazine*, Dec. 2018) raises a number of key issues at the intersection of technology and society. One of these is the challenge of "explainable AI."

For some reason the test pilots of new plane designs from WWII comes to my mind. I will first point out that most of these test pilots (in all of the countries) were women, because they wanted to fly, were not allowed in the fighting forces, and (deplorably) were considered expendable. While many could have learned much of the aerodynamics, and mechanical and other engineering aspects of each plane, this opportunity was not provided. So the reality is these pilots took to the air in a here-to-for untested aircraft on blind faith in the engineers and mechanics that put it on the runway. Needless to say, all of the players in the development team wanted each plane to work, but it is unlikely that any single person could explain it all, from rivet to gas gauge. Jump forward seventy years and ask if anyone can explain a Boeing 777.

Which leads to the second aspect of this, what happens when there is a failure? Boeing, and other manufacturers have analysis teams that evaluate all of the available information related to a failure, identify the most likely causes, and then identify the corrective action. This might entail new designs, additional equipment, reprogramming devices, or additional pilot training,

Similar attention needs to be applied to AI as it takes on an increasing role in our systems. In many ways, it already is embedded. We see this in cars, phones, television sets, and speakers that listen to what we say so that they can deliver the information or products we select. I'm not sure we are any better at explaining why we made a given request of our AI speakers than the speaker might be at explaining why it delivered the wrong thing.

But the challenge may be more ominous. The speaker may be directed to influence our decisions. "Nudging" is the term used in the IEEE standards work on Ethics for AI Design. An AI system that applies deep learning to manipulating human decisions, with detailed analysis of the targeted individual, is a disturbing potential that must affect our trust in both the systems and those that direct their applications.

As for explainable AI's, I must also give an example from Robert Heinlein's book, *The Moon is a Harsh Mistress*. When the voice enabled payroll system issues a check for one million dollars, the technician asks "why." The system answers, "I thought it would be funny."

Jim Isaak
CS2010@jimisaak.com

Digital Object Identifier 10.1109/MTS.2019.2913065
Date of publication: 30 May 2019