Hardwiring Hercules?

Courtney M. Cox

This Article seeks to reorient the debate over the right to a human decision. At present, the strongest arguments against machine decisionmakers, and in favor of humans, are grounded in explainability and transparency: Machines may now perform tasks once dismissed as "fantastical speculation," but only by using technology that is increasingly opaque and necessarily so. Thus, in contexts like adjudication where an explanation is owed, machines should not replace humans because only humans can provide the requisite explanation. Meanwhile, other principles advanced in favor of human decisionmakers have been deflated, reframed as grounding not a right to a human decision, but merely to a "better" decision—whether by human or machine.

This Article turns this debate on its head. First, it offers a reason to be skeptical of explainability arguments: Human judges—at least those who are not Herculean—experience what is called "normative uncertainty." But if they respond to that uncertainty rationally, they will have difficulty providing the kind of explanation demanded of machines. Whatever the merits of the right to an explanation, it would not seem to provide a reason to favor humans over machines in the quintessential example of decisions where an explanation is owed.

But what normative uncertainty gives with one hand, it takes with another. The problem of normative uncertainty limits the reach of the "better decision" argument in favor of machines and against the right to a human decisionmaker. The standard picture of automation, including most data-driven approaches, requires fixing in advance much of what, rationally, should be left open.

Recognizing and naming the problem provides language to diagnose that lingering worry—the pit of dread in one's stomach—that seems to remain even when other arguments in favor of machine decisionmakers are met or granted for argument's sake. It is cold comfort to say that we have a right to a "better" decision—and that AI can provide such a thing—when we remain uncertain about what "better" is. More critically, a failure to recognize this may lead both humans and machines astray in evaluating machine performance.