EXECUTIVE SUMMARY

How Humans Judge Machines

How Humans Judge Machines compares the reactions of people in the United States to scenarios describing human and machine actions.

Our data shows that people do not judge humans and machines equally, and that these differences can be explained as the result of two principles.

First, **people judge humans by their intentions and machines by their outcomes.**

By using statistical models to analyze dozens of experiments (chapter 6) we find that people judge machine actions primarily by their perceived harm, but judge human actions by the interaction between perceived harm and intention. This principle explains many of the differences observed in this book, as well as some earlier findings, such as people's preference for utilitarian morals in machines.

The second principle is that **people assign extreme intentions to humans and narrow intentions to machines.**

Technically, this means that people judge the intentions of humans using a bimodal distribution (either a lot or little intention) and the intention of machines using a unimodal distribution. This tells us that people are willing to excuse humans more than machines in accidental scenarios, but also that people excuse machines more in scenarios that can be perceived as intentional. This principle helps us explain a related finding—the idea that **people judge machines more harshly in accidental or fortuitous scenarios** (since they excuse humans more in such cases).

In addition to these principles, we find some specific effects. By decomposing scenarios in the five dimensions of moral psychology (harm, fairness, authority, loyalty, and purity), we find that **people tend to see the actions of machines as more harmful and immoral in scenarios involving physical harm.** Contrary to that, we find that **people tend to judge humans more harshly in scenarios involving a lack of fairness.** This last effect—but not the former—is explained mostly by differences in the intention attributed to humans and machines.

When it comes to labor displacement, we find that **people tend to react less negatively to displacement attributed to technology than to human sources**, such as offshoring, outsourcing, or the use of temporary foreign workers.

When it comes to delegation of responsibilities, we find that **delegating work to artificial intelligence tends to centralize responsibility up the chain of command.**

How Humans Judge Machines is a peer-reviewed academic publication. It was reviewed twice following the academic standards of MIT Press: once at the proposal stage (which included sample chapters), and again at full length. The experiments presented in this book were approved by the Internal Review Board (IRB) of the Massachusetts Institute of Technology (MIT).

These experiments involved 5,904 individuals who were assigned randomly to either a treatment (machine) or a control (human) group.

The scenarios in *How Humans Judge Machines* compare people's reactions to human and machine actions across the five dimensions of moral psychology, and visit contemporary issues such as algorithmic bias (chapter 3), privacy (chapter 4), and labor displacement (chapter 5).

We hope both humans and machines enjoy this book!

Sincerely,

César A. Hidalgo, PhD,
Artificial and Natural Intelligence Toulouse Institute (ANITI), University of Toulouse,
Alliance Manchester Business School, University of Manchester
School of Engineering and Applied Sciences, Harvard University