Kiel Brennan-Marquez is a professor of law and William T. Golden Scholar at UConn School of Law, where he teaches courses in constitutional law, evidence, and law and technology. His research focuses on the nature of human judgment and the role of courts—especially in the criminal context—in the age of big data and algorithmic decision-making.

Karen Levy is an assistant professor in the Department of Information Science at Cornell University, and associate member of the faculty of Cornell Law School. She researches how law and technology interact to regulate social life, with particular focus on social and organizational aspects of surveillance.

Daniel Susser is a philosopher by training, and works at the intersection of technology, ethics, and policy. He is an assistant professor in the College of Information Sciences and Technology, a research associate and core faculty member in the Rock Ethics Institute, and an affiliate faculty member in the Philosophy Department at Penn State University.

Abstract: The era of AI-based decision-making fast approaches, and anxiety is mounting about when, and why, we should keep “humans in the loop” (“HITL”). Thus far, commentary has focused primarily on two questions: whether, and when, keeping humans involved will improve the results of decision-making (making them safer or more accurate), and whether, and when, non-accuracy-related values—legitimacy, dignity, and so forth—are vindicated by the inclusion of humans in decision-making. Here, we take up a related but distinct question, which has eluded the scholarship thus far: does it matter if humans appear to be in the loop of decision-making, independent from whether they actually are? In other words, what is stake in the disjunction between whether humans in fact have ultimate authority over decision-making versus whether humans merely seem, from the outside, to have such authority? Our argument proceeds in four parts. First, we build our formal model, enriching the HITL question to include not only whether humans are actually in the loop of decision-making, but also whether they appear to be so. Second, we describe situations in which the actuality and appearance of HITL align: those that seem to involve human judgment and actually do, and those that seem automated and actually are. Third, we explore instances of misalignment: situations in which systems that seem to involve human judgment actually do not, and situations in which systems that hold themselves out as automated actually rely on humans operating “behind the curtain.” Fourth, we examine the normative issues that result from HITL misalignment.