STRANGE LOOPS: APPARENT VERSUS ACTUAL HUMAN INVOLVEMENT IN AUTOMATED DECISION MAKING

Kiel Brennan-Marquez,[†] *Karen Levy*[††] *& Daniel Susser*[†††]

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 at 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, arguing that it challenges individual decision-making about automated systems and complicates collective governance of automation.

DOI: https://doi.org/10.15779/Z385X25D2W

† Associate Professor of Law & William T. Golden Research Scholar, The University of Connecticut.

†† Assistant Professor of Information Science, Cornell University; Associated Faculty, Cornell Law School.

††† Assistant Professor of Information Sciences & Technology, and Research Associate in the Rock Ethics Institute, Penn State University. We thank Cassidy McGovern, Sherriff Balogun, and participants in the Cornell Tech Digital Life Initiative; the Cornell AI, Policy, and Practice Working Group; and the Berkeley Center for Law and Technology 2019 Symposium for helpful comments. Karen Levy gratefully acknowledges support from the John D. and Catherine T. MacArthur Foundation.

TABLE OF CONTENTS

I. INTRODUCTION

The era of automated decision making fast approaches, and anxiety is mounting about when and why we should keep "humans in the loop" (HITL).[1] Thus far, commentary has focused primarily on two questions: whether keeping humans involved will improve the results of decision making (rendering those results safer or more accurate),[2] and whether human involvement serves non-accuracy-related values like legitimacy and dignity.[3]

1. *See generally* Meg Leta Jones, *The Right to A Human in the Loop: Political Constructions of Computer Automation and Personhood*, 47 SOC. STUD. SCI. 216 (2017) (discussing the background on the burgeoning debate regarding whether to keep humans in the loop, particularly as it plays out in the United States-European Union context).

2. Medical treatment is a good example. Rich Caruana marshals a useful case study of asthmatic pneumonia patients who were categorized as "low risk" by a machine learning (ML) system—i.e., a system for automating classification tasks that infers or "learns" decision rules from prior examples rather than applying rules explicitly coded in advance—because it turns out that such patients (by contrast to non-asthmatic pneumonia patients) have historically received *much better care* from doctors, and so have displayed correspondingly better outcomes. In short, relying here on the ML system alone would have courted medical disaster. But the ML system was still a very useful input to ultimately-human decisions. *See* Rich Caruana et al., *Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission*, 21 ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING PROC. 1721, 1721–25 (2015).

3. *See, e.g.*, Kiel Brennan-Marquez & Stephen Henderson, *Artificial Intelligence and Role-Reversible Judgment*, 109 J. CRIM. L. & CRIMINOLOGY 137 (2019) (arguing that equality requires a "reversibility" dynamic between decision-makers and affected parties—and that this value

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 at 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?

Broadly speaking, our claim is that the "appearance" dimension of HITL merits exploration because when appearance and actuality are misaligned—when (1) a human appears to be in the loop, but in fact the decision-making system is fully automated, or when (2) a decision-making system appears fully automated, but is in fact bolstered by back-end human judgment—two related sets of normative issues come to the fore.

The first concerns individual experience. When appearance and actuality misalign, users of systems can become confused about what they are looking at. This dynamic risks both alienation and dignitary injuries, and deprives users of a meaningful opportunity to contest decisions.

The second set of normative issues attends to collective governance. Misalignment between the appearance and actuality of full automation can make it difficult to assess the ultimate goal of a decision-making system. Is full automation actually the desired endpoint? Are we—in the democratic, "we the people" sense—comfortable, in principle, with the automation of a given realm of decision making? Misalignment frustrates our ability to robustly ask these questions, regardless of their correct answers. Thus, where the stakes of automation are obscured by either a too-human or a falsely-inhuman veneer, democratic oversight suffers.

Our focus on the appearance of systems joins other recent legal scholarship focused on deceptive interfaces and the policy implications of humanrobot interaction.[4] Appearance emerges more latently in a good deal of other technology policy discussion. In fact, we might understand some of the most fundamental normative and policy principles in this area as efforts to align the actual and apparent operations of a system. Notice, for example, has long played a central role in policymaking around people's relationships with automated systems—most notably as a means of effectuating consent to

---

runs orthogonal to decisional outcomes); Emily Berman, *A Government of Laws and Not of Machines*, 98 B.U. L. REV. 1277 (2018) (arguing likewise); Frank Pasquale, *A Rule of Persons, Not Machines: The Limits of Legal Automation*, 87 GEO. WASH. L. REV. 1 (2019); Meg Leta Jones & Karen Levy, *Sporting Chances: Robot Referees and the Automation of Enforcement* (2017), https://ssrn.com/abstract=3293076 (pointing out the importance of sociocultural values like integrity and the overcoming of adversity in discussions of machine rule enforcement).

4. *See, e.g.*, Margot Kaminski et al., *Averting Robot Eyes*, 76 MD. L. REV. 983 (2017); Kate Darling, *'Who's Johnny?' Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy*, *in* ROBOT ETHICS 2.0 173 (Lin et al., eds., 2017).

data collection.[5] One of the most fundamental policy debates regarding the individualistic model of privacy regulation, and whether it can be resuscitated, involves the (in)effectiveness of privacy policies to provide notice that can serve as the basis for real consent.[6] The goal of notice, essentially, is to better align public perceptions with the actual workings of computational systems. Recent calls for interpretability of AI-driven systems, and explanations of the outcomes derived from them, have similar aims.[7]

Perhaps most fundamentally, appearances can help ensure the legitimacy of systems. Whether affected parties view decisions—particularly adverse decisions—as legitimate often depends on the presence of visible indicia of procedural regularity and fairness.[8] Sometimes, we go so far as to regulate these indicia regardless of the characteristics of the underlying system. In other words, sometimes we think appearances should be safeguarded, even if they make no difference to the ultimate decisions reached.[9] We require judicial recusal, for instance, both in cases where a judge is actually less-than-impartial, and in cases where it simply appears that way. The explicit justification for the latter—according to the American Bar Association and the Supreme Court—is that "appearance of impropriety" would "impair" the "*perception* [of a] judge's ability to carry out judicial responsibilities with integrity, impartiality and competence."[10] That is, it would threaten people's faith in the system, regardless of its impact on the case at hand.

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

5. *See generally* Daniel Susser, *Notice After Notice-and-Consent: Why Privacy Disclosures are Valuable Even If Consent Frameworks Aren't*, 9 J. INFO. POL'Y 37 (2019).

6. *See, e.g.*, Ryan Calo, *Against Notice Skepticism in Privacy (and Elsewhere)*, 87 NOTRE DAME L. REV. 1027 (2012) (exploring the concept of "visceral notice" as a means of revitalizing notice-and-consent regimes).

7. *See generally* Solon Barocas & Andrew Selbst, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085 (2018).

8. *See, e.g.*, TOM TYLER, WHY PEOPLE OBEY THE LAW (2006); John W. Meyer & Brian Rowan, *Institutionalized Organizations: Formal Structure as Myth and Ceremony*, 83 AM. J. SOC. 340 (1977).

9. *See* Roger Ford, Privacy When Form Does Not Follow Function (unpublished manuscript) (on file with author) (arguing that design changes can—profitably—impact the *experience* of user interaction with technology, even if they make no difference to actual technological capacity).

10. Caperton v. A.T. Massey Coal Co., 556 U.S. 868, 888 (2009) (emphasis added).

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.

II. HUMANS ACTUALLY IN THE LOOP VS. APPARENTLY IN THE LOOP

In recent years, the HITL question has become a focal point of technology-governance scholarship. This literature offers a handful of definitions of HITL. Some commentators construe HITL narrowly—to refer, in essence, to systems that operate automatically in the mine run of cases, but that provide for human override in circumstances of obvious error.[11] Other commentators define HITL more expansively—to encompass not only the possibility of case-by-case override by humans, but also the role of humans in developing and supporting automated systems, and the co-embeddedness of humans and machines in all technology-assisted decisional environments, "automated" or otherwise.[12] Although the observation that all technical systems are socially constructed certainly has conceptual value, the observation also makes it difficult to draw meaningful lines for present purposes.

In what follows, we deploy the concept of HITL to describe any decision-making system in which the initial triage or categorization of cases is performed by a machine, but a human agent exercises some degree of meaningful influence—up to and including override—over the disposition of particular cases. Influence takes different forms. Sometimes, the human role is largely procedural: for example, pushing a given case up or back in the relevant queue, or deciding which cases merit more institutional resources. Other times, the human role is more dispositive, involving the power to shape outcomes, either in terms of a case's concrete effects (e.g., granting or denying benefits), or in terms of how the outcome is justified, or both. The specifics of the human role may vary, but the key is that a human has some form of meaningful discretion in particular cases.[13]

11. For a formal model of HITL (specifically applied to security issues, but of general relevance) that goes in this direction, see Lorrie F. Cranor, *A Framework for Reasoning About The Human in the Loop*, 1 CONF. ON USABILITY, PSY., & SECURITY PROC. (2008).

12. *See, e.g.*, Meg Leta Ambrose, *The Law and the Loop*, 2014 IEEE INT'L SYMP. ON ETHICS SCI., TECH. & ENG'G 1 (2014) (emphasizing the universality of "humans in the loop," once the category is widened to include programmers, designers, and the like).

13. Our framing here tracks the conception of humans in the loop in the discourse around the European General Data Protection Regulation (GDPR), which triggers certain protections when decisions are made "based solely on automated processing"—that is, in the absence of a human in the loop. General Data Protection Regulation, 2016 O.J. (L 119)

Further, when we talk about "particular cases," we mean instances of decision making that have a concrete impact on a specific affected party—making the dynamic of interest the triangulated interaction of (1) the automated component of the system, (2) the HITL (who gets to decide, ultimately, what the fate of the affected party will be), and (3) the affected party herself. This is a capacious definition. As a formal category, it spans a diverse array of decision-making domains, some of which involve lots of "hands-on" human involvement, others of which involve almost none. Sometimes, the HITL and the affected party may be the same person, as in decision-making systems that empower—or *seem* to empower—users to directly override machine protocols. An especially pronounced and tragic example of this arose recently in two crashes of the Boeing 737 Max, despite pilots' efforts to override the software.[14] In both cases, one could say that the pilots were both the affected party of the machine-system and the HITL—or so, at least, it appeared.

At some level, however, the key point of our HITL definition is what it does not include. It does not include human involvement in the development of decision-making systems: the human aspects of coding, product design, or supervised learning. The reason is not that such human involvement lacks normative or practical relevance in these areas. It is that we are interested primarily in the impact of HITL—in actuality as well as appearance—on specific affected parties in decisional systems.

Our primary contribution is to add a dimension to the HITL discussion. Instead of simply asking whether a human *is* in the loop, we focus on whether a human *appears to be* in the loop. In other words, what has been traditionally conceptualized as a binary question—human in the loop: "yes" or "no"—may be better conceived as a 2x2 matrix. Enriching the model in

---

(EU) (repealing Directive 95/46/EC (General Data Protection Regulation)). In discussing the meaning of this provision, the Article 29 Working Party Guidelines maintain that "fabricating human involvement"—for instance, "if someone routinely applies [machine decisions] without any actual influence on the result"—would not escape the ambit of the automated processing provision. The report further clarifies that "[t]o qualify as human involvement, the controller must ensure that any oversight of the decision is meaningful, rather than just a token gesture. It should be carried out by someone who has the authority and competence to change the decision." ARTICLE 29 DATA PROTECTION WORKING PARTY, GUIDELINES ON AUTOMATED INDIVIDUAL DECISION MAKING AND PROFILING FOR THE PURPOSES OF REGULATION 2016/679, 20-21/en. wp 251rev.01 (Feb. 6, 2018) [hereinafter GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING].

14. Andrew J. Hawkins, *Deadly Boeing Crashes Raise Questions About Airplane Automation*, VERGE (Mar. 15, 2019, 1:40 PM), https://www.theverge.com/2019/3/15/18267365/boeing-737-max-8-crash-autopilot-automation [https://perma.cc/UU2L-GZ8Q].

this way—moving from a simple binary to a 2x2 matrix—helps us appreciate some of the normative complexity that attends the HITL debate.[15]

Table 1: HITL Dimensions

	Human is in the loop	**Human is not in the loop**
Human appears to be in the loop	I	II
Human does not appear to be in the loop	III	IV

On Table 1, quadrants I and IV are "aligned," meaning that the appearance of HITL and the actuality of HITL are the same. We call these quadrants *manifest humanity* and *full automation*. Quadrants II and III, by contrast, are "misaligned." Quadrant II, which we call *skeuomorphic humanity*, captures situations in which it seems like a human is present, but when a machine actually has full control. Think here of a chatbot with advanced language facility, or a home care robot that "seems human" to the patients for whom it cares. Inversely, quadrant III, which we call *faux automation*, captures situations in which the interface makes decision making seem completely automated, but where a human is actually making decisions—for example, a mobile robot that appears self-directed, but is in fact steered by a remote human driver. These definitions are included in Table 2.

15. To be sure, while our matrix adds a dimension to the HITL/no-HITL binary, it also necessarily collapses some real-life complexity. Just as a human may be more or less in the loop—that is, humans may have different degrees of discretion or autonomy vis-a-vis an automated system—the appearance of HITL is also not necessarily a binary. People may recognize, for instance, that a HITL is present, but misperceive the HITL's role. Or different users may be more or less recognizant of the true nature of the system. We elide such finer distinctions here for purposes of exploring the general dynamics, but recognize that they are likely to emerge in practice.

Table 2: HITL Dimensions with Definitions

	Human is in the loop	**Human is not in the loop**
Human appears to be in the loop	Manifest humanity	Skeuomorphic humanity
Human does not appear to be in the loop	Faux automation	Full automation

The reality, at least for the foreseeable future, is that many domains of automation will not be amenable to either of the two "aligned" quadrants. This is so for two reasons.

First, even in realms where total automation is plainly possible, the absence of humans in a process is likely to alienate some users. That is likely to inspire skeuomorphism, i.e., the *appearance* of human involvement. The companies and state agencies that develop automated technology, and the actors who deploy it, will have an incentive to use skeuomorphic techniques to drive adoption. Given this, it is plausible that many fully-automated realms will continue to maintain a veneer of human responsiveness. Techno-cultural evolution takes time.

Second, total automation will not be possible in certain realms for a long while. But it will nonetheless serve as an aspiration, and developers of technology will settle for faux automation as a bridge *toward* full automation. In other words, developers will often have an incentive to market systems which are not fully automated, on the promise—well-founded or not—that they will someday achieve full automation.

III. ALIGNMENTS

We begin with the two quadrants in which appearance and reality are consistent.

In the *manifest humanity* quadrant, a human is in fact in the loop, and this is apparent to users. Most forms of traditional adjudication fall within this category, as do uses of automated systems that serve purely to aid humans with well-established decision-making power (for example, the use of imaging technologies to assist doctors in medical diagnosis).

The inverse of manifest humanity is *full automation*—in which a process is completely and obviously automated with no human role. We may accept full automation as the best option when enforcement is low-stakes, uncontroversial, and rote—when an interest in efficiency outweighs other

normative concerns. At the other end of the spectrum, we may prefer fully automated systems in particularly high-stakes allocations of costs and benefits (like lotteries), in which we want no actual or apparent intervening value judgments about desert or blameworthiness.[16]

Each of these regimes may be advisable in some circumstances based on the values considerations we have discussed thus far (efficiency, fairness, safety, etc.). And both can be subject to legitimacy concerns on these or other grounds. We raise them here only in brief, primarily to set them aside. What interests us, ultimately, is the gap between appearance and reality—and its normative stakes.

IV. MISALIGNMENTS

A. SKEUOMORPHIC HUMANITY

Quadrant II encompasses cases of *skeuomorphic humanity*—situations in which the public generally perceives meaningful human involvement where none exists.

Human-like machine interfaces are ubiquitous. Sometimes, it is obvious to users that these machines are not actually human. Voice assistants like Siri and Alexa have notably human interactional qualities. They speak in humanoid voices, they tell jokes, and they respond to natural language queries. But their containment within a physical object like an iPhone or an Amazon Echo precludes most confusion that they are actually human. This is not always so. Online chatbots, for example, lack obvious indicia of their artificiality and often intentionally obscure it. They may do so for a variety of reasons, from efforts to deceive at scale (e.g., spambots and robocalls purporting to be from a human in need of a wire transfer) and economic and political manipulation (e.g., artificial generators of ratings and reviews; amplification of political propaganda) to therapeutic and even artistic goals (e.g., using bots to combat hate speech, or as a form of creative expression).[17] Google's artificial intelligence (AI) assistant Duplex—demonstrated at a May 2018 developer conference, in which it was used to book a haircut appointment—was purposefully given vocal qualities, tics and cadences that

16. This applies with particular force to intentionally randomized decisions. For example, Ronen Perry and Tal Zarsky discuss the attractiveness of purely random processes in high-stakes contexts like the law of the sea—if, say, one passenger must be thrown overboard to save the others, choosing the unlucky passenger by lot (presumably without subsequent appeal) may be the best way out of a bad situation. *See* Ronen Perry & Tal Z. Zarsky, *"May the Odds Be Ever in Your Favor": Lotteries in Law*, 66 ALA. L. REV. 1035, 1041 (2015).

17. *See* Madeline Lamo & Ryan Calo, *Regulating Bot Speech*, 66 UCLA L. REV. 988, 995–1002 (2019).

made it seem particularly realistic (pauses, "mm-hmm"s, and the like) to keep the person at the other end of the line from detecting its artificiality.[18]

In some cases, the skeuomorphic human is not a Siri-esque humanoid interface, but a real flesh-and-blood person—albeit one who lacks any meaningful ability to influence the relevant decision-making process. In these cases, the human is effectively no more than an ornamental aspect of the system's interface. These dynamics emerge in technical or bureaucratic systems that ostensibly involve humans, but where those humans are unable to execute discretion or diverge from administrative scripts. Think here about the familiar experience of visiting the DMV and being hamstrung by a minor technicality: for example, being told that one's insurance card needs to be in hard-copy rather than digital form in order to register a car, and that "no exceptions" can be made.[19] In practice, a human clerk is likely to deliver the news that one has failed to satisfy the agency's arcane requirements, suggesting that a well-reasoned or sufficiently emotional appeal might persuade them to revise the decision. But more often than not, the clerk merely throws up their hands and explains that they have no authority to override the rules. Although this decision-making system bears a human face, no human decision-maker impacts its outcomes (at least not in the immediate instance).

One defense of the "human gloss" is that it can make automated systems more intuitively usable. We borrow here from the vocabulary of *skeuomorphic design*—the use of design features that make an artifact resemble a previous version of itself.[20] In skeuomorphic design, the formerly functional becomes ornamental, a nod to prior technology that aids the user in transition.[21] For example, the "shutter click" sound of a phone camera: though the camera no longer has a physical shutter that makes such a sound, users have become

18. Interestingly, following blowback from critics about Duplex's deceptiveness, Google announced that a subsequent version would explicitly identify itself as an AI to the humans with whom it interacts. *See* Nick Statt, *Google Now Says Controversial AI Voice Calling System Will Identify Itself to Humans*, VERGE (May 10, 2018, 7:46 PM), https://www.theverge.com/2018/5/10/17342414/google-duplex-ai-assistant-voice-calling-identify-itself-update [https://perma.cc/5RFU-S876].

19. Readers who live in Connecticut be advised. In fairness to the state, DMV paperwork requirements were recently relaxed—registration applicants are now permitted to submit digital insurance cards. Though this, of course, does not make the system any more human; it simply makes the inhuman system more forgiving. An Act Concerning Electronic Proof of Automobile insurance identification cards, H.B. 5135, 2017 Sess. (Conn. 2009), https://www.cga.ct.gov/asp/cgabillstatus/cgabillstatus.asp?selBillType=Bill&which_year=2017&bill_num=5135 [https://perma.cc/C5EE-8NF4].

20. *Skeuomorphism*, INTERACTION DESIGN FOUND., https://www.interaction-design.org/literature/topics/skeuomorphism [https://perma.cc/29WA-JUXU] (last visited Oct. 18, 2019).

21. *See id.*

acclimated to the idea that shutter click indicates a photo taken. Therefore, subsequent technologies have included the sound as an ornament. The ornament retains social functionality by acting as a signifier, a notification to photo takers and photo subjects that a photo has been captured.[22] (Think, too, of e-readers with "pages," or digital audio controls shaped like dials.[23])

We might think of skeuomorphism as a form of *design theater*.[24] Interaction with artifacts and processes often involves a sort of ritualism; our understanding of technologies depends on how we have interacted with them in the past. When something about the technology changes in a way that obviates that ritual, we may be put off or confused. The retention of ritual—even when not strictly necessary for the system to function technically—can help the system to function socially. Consider, for instance, the legend of midcentury cake mix.[25] As the story goes, home cake mixes—in which all ingredients save water were pre-measured and mixed together, so that the baker need only dump the box's contents into water, stir, and bake—initially sold poorly. Psychologist Ernest Dichter recommended that General Mills reformulate the mix to require more human work. The reason, Dichter offered, was that housewives found the process self-indulgent: "In order to enjoy the emotional rewards of presenting a homemade cake, they had to be persuaded that they had really baked it, and such an illusion was impossible to maintain if they did virtually nothing."[26] As a result, it is said, the company changed the recipe to require that the baker add fresh eggs to the mix in place of the dehydrated eggs that had been included. This change ostensibly led to the product's wide acceptance. The story suggests that even when not essential for technical functioning, the patina of humanity in a process can matter.

Further, even in realms where we are comfortable with full automation as a normative matter—i.e., the decision-making task is not one that seems,

22. *See* John H. Blitz, *Skeuomorphs, Pottery, and Technological Change*, 117 AM. ANTHROPOLOGIST 665, 668 (2015) (describing skeuomorphs as both "utilitarian and representational"); *see also* Ivan Marković, *Vaping Like a Chimney: Skeuomorphic Assemblages and Post-Smoking Geographies*, SOC. & CULTURAL GEOGRAPHY 1, 2 (2019) (presenting a conceptual overview of the skeuomorph).

23. *See* Tim Hwang & Karen Levy, *The Presentation of Machine in Everyday Life*, WEROBOT (Mar. 2015), http://www.werobot2015.org/wp-content/uploads/2015/04/Hwang_Levy_WeRobot_2015.pdf [https://perma.cc/K6CU-TLYR].

24. *Id.*

25. The minutiae of the story itself are contested, and possibly apocryphal, but it serves its purpose here regardless. *See* David Mikkelson, *Requiring an Egg Made Instant Cake Mixes Sell?*, SNOPES (Jan. 31, 2008), https://www.snopes.com/fact-check/something-eggstra/ [https://perma.cc/8EAC-BGZL].

26. PAUL LEE TAN, ENCYCLOPEDIA OF 7700 ILLUSTRATIONS: SIGNS OF THE TIMES 1228 (1979).

in principle, to require human judgment—there may be still be dignitary reasons to maintain the appearance of humanity, even in a purely ministerial capacity. A good example is the delivery of momentous information, as in recent debates over whether doctors should deliver grave prognoses via robot.[27] Many people think that dire medical information deserves some kind of "cushion," or human gloss, which might be a freestanding argument for keeping the skeuomorphic structure in place.[28] It is also possible for the appearance of human involvement to help smoothly transition a decision-making system to full automation. This is not an argument in favor of maintaining skeuomorphic structures perpetually, but can certainly justify maintaining them in the short- to medium-term.[29] Acknowledging these benefits is quite different from wanting a human to *actually* be meaningfully involved in decision making.[30] The objection here is not to the means of arriving at the prediction, but to the method by which that prediction is communicated.

Design of this sort is not without detractors. Although some preferences are purely aesthetic, others depend on design theater's tendency to enable deception or manipulation, when users are made to feel comfortable with a new technology because they think it works just like an older one.[31] Often, design theaters operate to give users the feeling of being in greater control over a technology than they actually are (what we have elsewhere called

27. *See* David Aaro, *Family Upset After 'Robot' Doctor Informs Patient He Doesn't Have Long to Live*, FOX NEWS (Mar. 10, 2019), https://www.foxnews.com/health/family-upset-after-robot-doctor-says-patient-doesnt-have-long-to-live [https://perma.cc/M6WH-UYT7] (" 'If you're coming to tell us normal news, that's fine, but if you're coming to tell us there's no lung left and we want to put you on a morphine drip until you die, it should be done by a human being and not a machine,' Catherine Quintana told USA Today."); Evan Selinger & Arthur Caplan, *How Physicians Should and Shouldn't Talk with Dying Patients*, ONEZERO (Mar. 12, 2019), https://onezero.medium.com/how-physicians-should-and-shouldnt-talk-with-dying-patients-6ff55fcf40e4 [https://perma.cc/C39F-NS89]; Joel Zivot, *In Defense of Telling Patients They're Dying via Robot*, SLATE (Mar. 13, 2019), https://slate.com/technology/2019/03/robot-doctor-technology-patient-dying.html [https://perma.cc/8CFT-R3PP]. Notably, the human doctor did appear on the robot's screen and delivered the news via videoconference—but the means of communication nevertheless caused injury and offense.

28. *See* Zivot, *supra* note 27.

29. Katherine Metcalf et al., *Mirroring to Build Trust in Digital Assistants*, ARXIV (Apr. 2, 2019), https://arxiv.org/pdf/1904.01664.pdf [https://perma.cc/3NBJ-A8DF].

30. Brennan-Marquez & Henderson, *supra* note 3.

31. A somewhat comical, but instructive example is the "Horsey Horseless," a turn-of-the-19th-century vehicle design that consisted, essentially, of "a car with a big wooden horse head stuck on the front of it," intended to mislead *horses* on the road into accepting a motorized vehicle as one of their own. It does not appear to have worked. Alex Davies, *Well That Didn't Work: The 1899 Car With a Full-Size Wooden Horse Head Stuck to the Front*, WIRED (Feb. 10, 2015), http://www.wired.com/2015/02/well-didnt-work-1899-car-full-size-wooden-horse-head-stuck-front/ [https://perma.cc/ZU57-GFD5].

"theaters of volition")—like placebo buttons that give users the illusion of agency over elevator doors or crosswalk signals.[32] Speed is another common consideration: users may not trust computational processes that occur instantaneously, so designers may deliberately build delay and the appearance of deliberation or processing into systems.[33] Cases like these deceive users by deliberately obscuring the full capabilities of the system and the limited abilities of the human user.

Sometimes the concern is less about deception than visceral aversion. Human-like machines launch us into the uncanny valley—things that look almost, but not quite, like humans make us feel very uncomfortable.[34] There are several different explanations for this feeling of eeriness. One cognitive explanation is that when it is harder for us to categorize something immediately, we have a sense of dissonance and discomfort that is difficult to resolve. An explanation from evolutionary psychology is that vaguely unnatural movement can be an indicator of pathogens, so we are conditioned to want to stay away from it.[35] Regardless of the source, being duped by a machine masquerading as a human is an uncomfortable feeling.

More pragmatic concerns attach, too. Human-seeming systems can readily gain our trust—or manipulate us, leading to a range of consumer protection issues.[36] We may disclose more to human-seeming systems than we otherwise might, perhaps because we have misread human-like cues.[37] The mistaken sense that a human is involved in an automated process can lead people to believe that there are more opportunities for intervention and override than actually exist. Ultimately—as we explore more fully in Part IV below—the key question is whether maintaining the appearance of human involvement has sufficient benefits to outweigh the inherent shortcomings of deception.[38]

---

32. Hwang & Levy, *supra* note 23; Torin Monahan, *Built to Lie: Investigating Technologies of Deception, Surveillance, and Control*, 32 INFO. SOC'Y 229 (2016).

33. *See* Ryan W. Buell & Michael I. Norton, *The Labor Illusion: How Operational Transparency Increases Perceived Value*, 57 MGMT. SCI. 1564 (2011).

34. *See* Shensheng Weng et al., *The Uncanny Valley: Existence and Explanations*, 19 REV. GEN. PSYCHOL. 393 (2015).

35. Karl F. MacDorman et al., *Too Real for Comfort? Uncanny Responses to Computer Generated Faces*, 25 COMPUTERS HUM. BEHAV. 695, 696 (2009).

36. *See generally* Woodrow Hartzog, *Unfair and Deceptive Robots*, 74 MD. L. REV. 785 (2015).

37. Brenda Leong & Evan Selinger, *Robot Eyes Wide Shut: Understanding Dishonest Anthropomorphism*, 19 ACM FAT CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 299, 299 (2019).

38. *See generally* Eytan Adar et al., *Benevolent Deception in Human Computer Interaction*, 13 ACM CONF. ON HUMAN COMPUTER INTERACTION (CHI) (2013) (providing a thorough description of rationales and methods for user deception in human-computer interaction).

B. FAUX AUTOMATION

Quadrant III points to the inverse of skeuomorphic humanity—what is sometimes called *faux automation* (or what writer and activist Astra Taylor calls *fauxtomation*).[39] Here, the misalignment between appearance and reality arises because apparently automated systems are in fact driven by considerable human input. Of course, as scholars in science and technology studies (STS) have long argued, at some level, all technologies reflect the concerns, perspectives, and values of their human designers.[40] By *faux automation*, however, we suggest more direct forms of human involvement, consistent with our definition of HITL above.

The reason for faux automation is straightforward: building fully automated systems is hard. Despite recent advances in machine learning and AI, certain tasks that humans easily accomplish, such as understanding and using words in context, remain difficult for computers.[41] Rather than wait for further breakthroughs, technologists increasingly conceive of automation problems outside binary, all-or-nothing terms (full automation or bust), and use hybrid human-machine workflows to solve complex problems. Amazon's Mechanical Turk (AMT) system, a major platform for coordinating such work, originally described itself as facilitating "artificial artificial intelligence": a simulacrum of automation, in which humans masquerade as machines that think like humans.[42]

Examples of faux automation abound, exhibiting a variety of human-machine configurations. In some arrangements, machines do most of the work and human involvement is largely limited to quality assurance. For example, it was recently revealed that Amazon's Alexa devices—voice-activated "smart assistants" advertised as using AI to answer users' questions

39. Astra Taylor, *The Automation Charade*, LOGIC (Aug. 1, 2018), https://logicmag.io/05-the-automation-charade/ [https://perma.cc/2YCJ-2MCM].

40. *See, e.g.*, Ambrose, *supra* note 12; *see generally* BATYA FRIEDMAN & DAVID G. HENDRY, VALUE-SENSITIVE DESIGN: SHAPING TECHNOLOGY WITH MORAL IMAGINATION (2019).

41. Will Knight, *AI's Language Problem*, MIT TECH. REV. (Aug. 9, 2016), https://www.technologyreview.com/s/602094/ais-language-problem [https://perma.cc/38ZZ-XJKZ].

42. Using AMT, "requestors" distribute small work assignments ("Human Intelligence Tasks," or HITs, as Amazon calls them)—e.g., identifying objects in an image or digitizing handwritten text—to a distributed, online workforce ("turkers"), who are paid per task completed. *Artificial Artificial Intelligence*, ECONOMIST (Jun. 10, 2006), https://www.economist.com/technology-quarterly/2006/06/10/artificial-artificial-intelligence?story_id=7001738 [https://perma.cc/N4KT-FW5B].

and to control other "smart home" systems—fall into this category.[43] Unbeknownst to Alexa owners, who were given the impression that the devices are fully automated, audio recordings of user prompts and queries are regularly transmitted back to Amazon, where human technicians review them in order to tweak and improve Alexa's algorithms.[44]

At the other end of the spectrum are cases in which humans do most of the thinking and machine components are largely for show. One example is the *original* Mechanical Turk, an 18th century chess-playing automaton that turned out to have a human chess player hidden inside its enclosure. These systems are designed to give the appearance of automation without the computational substance.[45] In 2015, the public learned that the Edison automated blood testing systems sold by Silicon Valley firm Theranos were just this kind of charade.[46] Theranos advertised its Edison machines as a revolutionary technology that could process hundreds of diagnostic tests using only a few drops of blood instead of the numerous vials older techniques required. But the machines did not work.[47] Rather than admit it, the company staged misleading demonstrations and falsified Food and Drug Administration tests. The company pretended that its own machines were processing the blood, when lab technicians were actually conducting the tests behind the scenes using standard industry equipment purchased from their competitors.[48]

Many faux automated systems rely on human-machine collaborations that fall somewhere between these extremes. While significant functionality is automated, humans are generally responsible for tasks such as text and image recognition. In 2017, it came to light that Expensify (an app for generating expense reports) was using human workers contracted through AMT to digitize handwritten receipts.[49] In 2018, the Center for Public Integrity exposed widespread errors in campaign finance records caused by human mislabeling of images being prepared for automated processing by a

43. Matt Day et al., *Amazon Workers Are Listening to What You Tell Alexa*, BLOOMBERG (Apr. 10, 2019), https://www.bloomberg.com/news/articles/2019-04-10/is-anyone-listening-to-you-on-alexa-a-global-team-reviews-audio [https://perma.cc/92KW-M4AD].

44. *Id.*

45. Taylor, *supra* note 39.

46. John Carreyrou, *Hot Startup Theranos Has Struggled With Its Blood-Test Technology*, WALL ST. J. (Oct. 16, 2015), https://www.wsj.com/articles/theranos-has-struggled-with-blood-tests-1444881901 [https://perma.cc/K6G7-WWLU].

47. *Id.*

48. *Id.*

49. Alison Griswold, *Expensify's "Smart" Scanning Technology Was Secretly Aided by Humans*, QUARTZ (Nov. 30, 2017), https://qz.com/1141695/startup-expensifys-smart-scanning-technology-used-humans- hired-on-amazon-mechanical-turk/ [https://perma.cc/4WR6-9TKA].

company called Captricity.[50] Even more difficult for machines than text and image recognition is judging the meaning of words and images in context. This makes human workers essential to commercial content moderation.[51] While Facebook CEO Mark Zuckerberg has promised that the company's AI tools will rid its platform of problematic content, there is little reason to believe fully automated content moderation systems are on the horizon.[52] Facebook and other social media sites, such as Twitter and YouTube, have devised elaborate rules for determining when user-generated content should be flagged or removed—systems that Kate Klonick has likened to "legal or governance systems."[53] But in many cases, machines are incapable of determining which rules apply to particular posts, or deciding when the rules need to be revised or amended.[54] Thus, armies of human reviewers are required to carry out this interpretive work.[55]

Humans may also play a significant role in seemingly autonomous robotic systems. The Kiwibot, a four-wheeled food delivery robot currently deployed for testing on the UC Berkeley campus, is actually operated by workers in Colombia who send the robots wayfinding instructions every five to ten seconds. (The arrangement, which the company calls "parallel autonomy," saves money because the humans obviate the need for sophisticated sensor systems).[56] Similarly, a Japanese firm called Mira Robotics recently announced the release of remote-controlled "robot butlers" (think Rosie from *The Jetsons*). These robots rely on a combination of AI software for basic navigation and remote human controllers for more

50. Rosie Cima, *Company Using Foreign Workers Botches U.S. Senate Campaign Finance Records*, CTR. FOR PUBLIC INTEGRITY (Sep. 5, 2018), https://publicintegrity.org/federal-politics/company-using-foreign-workers-botches-u-s-senate-campaign-finance-records/ [https://perma.cc/G6XP-G7EN].

51. Sarah T. Roberts, *Social Media's Silent Filter*, ATLANTIC (Mar. 8, 2017), https://www.theatlantic.com/technology/archive/2017/03/commercial-content-moderation/518796/ [https://perma.cc/W9K4-PAV9] ("[T]here is a profound human aspect to this work.").

52. Drew Harwell, *AI Will Solve Facebook's Most Vexing Problems, Mark Zuckerberg Says. Just Don't Ask When or How.*, WASH. POST (Apr. 11, 2018), https://www.washingtonpost.com/news/the-switch/wp/2018/04/11/ai-will-solve-facebooks-most-vexing-problems-mark-zuckerberg-says-just-dont-ask-when-or-how/ [https://perma.cc/G2FJ-2CBG].

53. *See* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1602 (2018).

54. *Id.* at 1635–49.

55. *Id.*

56. Carolyn Said, *Kiwibots Win Fans at UC Berkeley As They Deliver Fast Food at Slow Speeds*, S.F. CHRON. (May 26, 2019), https://www.sfchronicle.com/business/article/Kiwibots-win-fans-at-UC-Berkeley-as-they-deliver-13895867.php [https://perma.cc/583D-WXDC].

complex tasks like folding clothes and manipulating small objects.[57] Although Mira Robotics has been forthright about its robots' human control, as these kinds of devices proliferate we can expect the gap between user perceptions about the nature of these systems and the reality of their internal functioning to grow.

Faux automation and skeuomorphic humanity are not mutually exclusive: one system can exhibit both dynamics. Consider Google's Duplex service, previously described. Originally debuted as "a new technology for conducting natural conversations to carry out 'real world' tasks over the phone," the system was designed as an outward-facing AI assistant.[58] Rather than merely answer questions, it could call and schedule reservations and appointments, speaking to other people on its user's behalf.[59] In Google's initial demonstrations, Duplex did not disclose to the people it called that they were speaking to a machine—a case of skeuomorphic humanity—and skeptics quickly raised alarms about the deception involved.[60] But a more complex revelation followed: Duplex's algorithms required significant human help in order to function. Confronted by the New York Times, Google admitted that "about 25 percent of calls placed through Duplex started with a human, and that about 15 percent of those that began with an automated system had a human intervene at some point."[61] Faux automation was thus used as a stop-gap on the way to skeuomorphic humanity—a human pretending to be a machine, while the machine pretended to be a human.

The illusion of automation gives rise to at least two distinct concerns. First, there may be contexts in which we would welcome machine assistance, but balk at human help. Smart speakers are designed to record us in what was

57. James Vincent, *Robot Butlers Operated by Remote Workers are Coming to Do Your Chores*, VERGE (May 9, 2019), https://www.theverge.com/2019/5/9/18538020/home-robot-butler-telepresence-ugo-mira-robotics [https://perma.cc/9AJY-343T].

58. Yaniv Leviathan & Yossi Matias, *Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone*, GOOGLE AI BLOG (May 8, 2018), https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html [https://perma.cc/6BTD-ABVZ].

59. *Id.*

60. *See, e.g.*, Brian Feldman, *Google Duplex Makes Your Life Easier by Making It More Difficulty for Others*, N.Y. MAG. (May 10, 2018), http://nymag.com/intelligencer/2018/05/google-duplex-no-no-no-no-no-no.html [https://perma.cc/P5RN-9RPK]; Alex Hern, *Google's 'Deceitful' AI Assistant to Identify Itself as a Robot During Calls*, GUARDIAN (May 11, 2018), https://www.theguardian.com/technology/2018/may/11/google-duplex-ai-identify-itself-as-robot-during-calls [https://perma.cc/6W42-3F6K]; Natasha Lomas, *Duplex Shows Google Failing at Ethical and Creative AI Design*, TECHCRUNCH (May 10, 2018), https://techcrunch.com/2018/05/10/duplex-shows-google-failing-at-ethical-and-creative-ai-design/[https://perma.cc/F2HV-J48C].

61. Brian X. Chen & Cade Metz, *Google's Duplex Uses A.I. to Mimic Humans (Sometimes)*, N.Y. TIMES (May 22, 2019), https://www.nytimes.com/2019/05/22/technology/personaltech/ai-google-duplex.html [https://perma.cc/H9WJ-9DFL].

once called the privacy of our own homes, and Amazon markets some of its Alexa devices, such as the "Echo Spot" smart alarm clock, for installation in the bedroom.[62] Yet those comfortable with having their intimate conversations monitored by Amazon's algorithms may feel differently about having them heard by human listeners.[63] This concern has also been raised in relation to robots: if people are "deceived into thinking the robot is acting autonomously" rather than being human-controlled, they may "disclose sensitive information to the robot that they would not tell a human, not realizing that a human is hearing everything they say."[64] This was equally true in the Expensify case, discussed above. Expensify users, under the impression that machines were digitizing their receipts, were dismayed to learn that human AMT workers read and transcribed them, as receipts often contain sensitive personal information.[65]

Second, the appearance of automation can disguise the mistreatment of human workers behind the scenes.[66] Work managed through AMT is not typically well-paid. While Amazon does not provide precise wage figures, estimates suggest that "turkers" (i.e., AMT workers) earn on average only $2 per hour.[67] In addition to wage issues, the nature of the work can be distressing and damaging. Researchers and journalists have chronicled the gruesome text, images, and videos that commercial content moderators must endure in order to purge such content from our social media feeds, and the inadequate support tech companies often provide them.[68] Yet much of this

62. Tom Warren, *Amazon's Echo Spot is a Sneaky Way to Get a Camera Into Your Bedroom*, VERGE (Sep. 28, 2017), https://www.theverge.com/2017/9/28/16378472/amazons-echo-spot-camera-in-your-bedroom [https://perma.cc/W26A-PZWY].

63. Hartzog, *supra* note 36, at 794.

64. Jacqueline Kory Westlund & Cynthia Breazeal, *Deception, Secrets, Children, and Robots: What's Acceptable?*, HUM. ROBOT INTERACTION WORKSHOPS (2015).

65. Griswold, *supra* note 49.

66. Taylor, *supra* note 39.

67. Kotaro Hara et al., *A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk*, 18 ACM CONF. ON HUM. FACTORS IN COMPUTING SYS. 11 (2018) ("We estimate that 96% of workers on AMT earn below the U.S federal minimum wage. While requesters are paying $11.58/h on average, dominant requesters who post many low-wage HITs like content creation tasks are pulling down the overall wage distribution."). Additionally, Kiwibot operators also make less than $2 per hour. Said, *supra* note 56.

68. Sarah T. Roberts, *Social Media's Silent Filter*, ATLANTIC (Mar. 8, 2017), https://www.theatlantic.com/technology/archive/2017/03/commercial-content-moderation/518796/ [https://perma.cc/NDS5-9ZQE]; Sarah T. Roberts, *Meet the People Who Scar Themselves to Clean Up Our Social Media Networks*, MACLEAN'S (Jun. 15, 2018), https://www.macleans.ca/opinion/meet-the-people-who-scar-themselves-to-clean-up-our-social-media-networks/ [https://perma.cc/R6V7-DHEP]; Adrian Chen, *The Human Toll of Protecting the Internet from the Worst of Humanity*, NEW YORKER (Jan. 28, 2017), https://www.newyorker.com/tech/annals-of-technology/the-human-toll-of-protecting-the-internet-from-the-worst-of-humanity [https://perma.cc/PYP4-29NU].

work is rendered invisible, because users are led to believe that these systems are fully automated.[69]

V. HOW MISALIGNMENT UNDERMINES REASONING ABOUT AUTOMATION

The misalignments described in the previous section provoke normative worry both at the individual and institutional levels. Beneath both sets of problems lies the same fundamental issue: misalignment sows confusion. It undermines our capacity to understand and reason about automated systems. For individuals, misalignment makes it difficult to contest or resist the decisions these systems deliver. For institutions, misalignment frustrates governance; it hinders the public's ability to discern and meaningfully balance the benefits and harms of automation.

These problems manifest differently in cases of skeuomorphic humanity and in cases of faux automation. In cases involving skeuomorphic humanity, individuals confronting human-seeming, but in fact fully automated systems have no real opportunity for appeal. The human acts as a bait-and-switch, palliating users' concerns without offering real recourse. Consider again the case of a DMV agent who refuses to deviate from their administrative script, even when the decision it reaches is arguably unreasonable. Set at ease by a human veneer, we expect that a human—with the apparent power to intervene or override the system's rote determination—will hear our grievances. Instead we find that resistance is futile.[70]

In cases of faux automation, by contrast, misalignment misdirects, rather than thwarts, our attempts at contesting the system's judgments. For example, if users are given the impression that content moderation on a social media platform has been fully automated, when in fact it is carried out in large part by an army of human reviewers, they are misled about the

69. *See generally* MARY GRAY & SIDDHARTH SURI, GHOST WORK: HOW TO STOP SILICON VALLEY FROM BUILDING A NEW GLOBAL UNDERCLASS (2019).

70. Ben Wagner points out that apparent-but-not-actual HITL (what he terms *quasi-automation*, and what we call skeuomorphic humanity) can frustrate the aims of legal rules, as well. Laws that aim to promote human rights with respect to algorithmic decision-making (notably, the GDPR) assume that HITLs have some measure of agency and influence; if they do not, they amount to no more than "a human fig-leaf for automated decisions" that cannot adequately safeguard rights. Ben Wagner, *Liable, But Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems*, 11 POL'Y & INTERNET 104, 118 (2019). Wagner proposes seven criteria through which to define when a human is meaningfully in the loop, as opposed to when one is simply present to "rubber-stamp" automated decisions. *Id.* at 115.

source of problems.[71] Rather than focusing indignation on the human process that caused the mistake, people tend to lodge their grievances against automation. This result verges on ironic, since genuine automation may well be a solution to the problem (depending on our diagnosis of what the problem is), rather than its cause.

Similar issues arise at the collective or institutional level. To the extent that decision-making systems are performing sub-optimally, misalignment distorts our impression of the problem. Specifically, misalignment between the appearance and reality of human control over decision making can cause certain normative dynamics to become ambiguous or insufficiently differentiated. This is unfortunate for at least two reasons. First, different dynamics, once identified, raise different governance issues. Ambiguities between dynamics therefore produce a risk of solutions that poorly fit, or even disserve, the problem at hand. Second, the question of what dynamic we are confronting—the nature of the problem—will often be a source of normative controversy in its own right. In other words, there are many circumstances in which no "right answer" exists to the question of which dynamic is afoot. Rather, the issue is essentially and irreducibly political, such that even the question of how to conceptualize the problem calls out for democratic oversight.

To get a better sense of what we mean, consider each of the following dynamics, grouped according to which form of misalignment—skeuomorphic humanity or faux automation—they reflect. In each, we consider how normative issues can emerge based on the ideal calibration, in terms of the appearance and actuality of human involvement, for a given decision-making system.

A. DYNAMICS RELATED TO SKEUOMORPHIC HUMANITY

1. The first dynamic is that, ideally, a decision-making system would both be and appear automated—but at present it appears non-automated.

71. *See* James Vincent, *AI Won't Relieve the Misery of Facebook's Human Moderators*, VERGE (Feb. 17, 2019), https://www.theverge.com/2019/2/27/18242724/facebook-moderation-ai-artificial-intelligence-platforms [https://perma.cc/2VPX-E5VQ].

Table 3: Unrealized Ideal—Full Automation

	Human is in the loop	**Human is not in the loop**
Human appears to be in the loop		PRESENT STATUS QUO (skeuomorphic humanity)
Human does not appear to be in the loop		UNREALIZED IDEAL (full automation)

Here, the skeuomorphic quadrant is essentially an interim position: the problem is not that the decision-making system *is* insufficiently automated, but that it *looks* insufficiently automated. And once again the key governance issue becomes whether it is possible—and desirable—to move toward a greater appearance, or awareness, of automation. When the answer is yes, the practical question becomes how best to facilitate the transition: by what means, on what timetable, at whose cost, and the like. Proposed chatbot disclosure laws are a good example of an effort to move in this direction.[72] By requiring overt disclosure of the machine nature of a chatbot, the user is presumably not deceived into believing she is communicating with a human, and can modulate her behavior accordingly.[73]

2. The second dynamic is that, ideally, a decision-making system would neither appear to be, nor actually be fully automated, but at present it *is* automated.

72. *See, e.g.*, Jeffrey D. Neuburger & Daryn A. Grossman, *Get All of Your Bots in a Row: 2018 California Bot Disclosure Law Comes Online Soon*, NAT'L L. REV. (June 7, 2019), https://www.natlawreview.com/article/get-all-your-bots-row-2018-california-bot-disclosure-law-comes-online-soon [https://perma.cc/98X8-6M72].

73. *But see* Lamo & Calo, *supra* note 17, at 6 (noting that even if bots are revealed as bots, they "can [still] cause harm, primarily by tricking and confusing consumers. Robocallers may deny that they are automated, call targeted individuals repeatedly, and even claim to be a representative of the IRS or another powerful entity that even a tech-savvy individual might feel too anxious to hang up on").

Table 4: Unrealized Ideal—Manifest Humanity

	Human is in the loop	**Human is not in the loop**
Human appears to be in the loop	UNREALIZED IDEAL (manifest humanity) ←	PRESENT STATUS QUO (skeuomorphic humanity)
Human does not appear to be in the loop		

The governance questions on this front are straightforward in theory, but often complex in practice. In principle, the issue is simply one of putting a human "back" into the loop—a reversion to the pre-automated world. But in practice, at least two wrinkles emerge. The first is that reversion is often costly, and directly contrary to the economic interests of the actors, governmental or corporate, who spearheaded the effort toward automation in the first place. So, at a minimum, significant political will is required. The second wrinkle is that even those who agree about the need to reinsert a human in the loop will likely dispute *how* to do so. At what point(s) in the process should human oversight be installed? And what kind of oversight? And—as ever—which humans? These issues may emerge particularly when the combination of automation and deception removes some socially important friction. For instance, while bot disclosure laws require a change in the *appearance* of a chatbot, proposed anti-robocall legislation takes a different tack by banning certain types of automated calling altogether.[74] Doing so makes direct marketing much costlier for companies making these calls, and presumably realigns their incentives to do so.

3. The third dynamic is that, ideally, a decision-making system would be automated, but not seem so, making the skeuomorphic quadrant not simply an interim state, but a direct realization of the ideal.

74. Emily Birnbaum, *Dem Chair Offers Bill to Crack Down on Robocalls*, HILL (Feb. 4, 2019), https://thehill.com/policy/technology/428372-dem-introduces-bill-to-crack-down-on-robocalls [https://perma.cc/X399-UDMU].

Table 5: Realized Ideal

	Human is in the loop	**Human is not in the loop**
Human appears to be in the loop		PRESENT STATUS QUO & REALIZED IDEAL (skeuomorphic humanity)
Human does not appear to be in the loop		

A good illustration of this dynamic is a care-bot that assists ill and elderly people.[75] Assuming for argument's sake that at least some care functions are susceptible to automation, it does not follow that "full automation" is the ideal paradigm. For it may be that other, countervailing considerations—for example, the psychological benefits that come from being cared for in a human-feeling way—may counsel in favor of continued, even perpetual, skeuomorphism. Indeed, this is precisely why many skeuomorphs exist: they lubricate the transition from Technological Environment A to Technological Environment B for the human subjects who occupy, and interact within, those environments. Sometimes, this process is self-consciously temporary. Other times, it can be indefinite, particularly when the skeuomorph evolves into a comfortable feature of Technological Environment B, despite its lack of functional purpose. Think, for instance, of the persistent use of "buttons" in UX design. There is no functional reason that screen-based interfaces must include button-shaped mechanisms of navigation. Yet people seem to like them, and understand how to use them, and it is therefore conceivable that they will persist for a long time to come.

Yet even in this case—despite the status quo overlapping formally with the ideal—many second-order governance questions remain. What are the goals of the skeuomorphic mechanism and how do they potentially trade off against other goals? Having answered that question to satisfaction, what are the specific design features of the skeuomorphic mechanism that best balance these goals?

B. DYNAMICS RELATED TO FAUX AUTOMATION

The possible dynamics with respect to faux automation form a mirror-image of those just explored.

75. Don Lee, *Desperate for Workers, Aging Japan Turns to Robots for Health Care*, SEATTLE TIMES (Jul. 30, 2019), https://www.seattletimes.com/business/desperate-for-workers-aging-japan-turns-to-robots-for-health-care/ [https://perma.cc/X5FL-NVMM].

1. The first dynamic is that, ideally, a decision-making system would both seem and be fully automated—but at present only seems automated, without actually being so.

Table 6: Unrealized Ideal—Full Automation

	Human is in the loop	**Human is not in the loop**
Human appears to be in the loop		
Human does not appear to be in the loop	PRESENT STATUS QUO (faux automation) →	UNREALIZED IDEAL (full automation)

This gives rise to two interrelated governance questions: (1) whether it is possible or realistic, given existing technology, to move toward actual automation, and (2) what the drawbacks of doing so would be. In other words, as with the equivalent dynamic above, here the faux automation quadrant is an interim state. Although full automation is the ideal, the status quo involves faux automation—and the question becomes whether it is possible (and, all things considered, desirable) to move toward the former.

Certain compliance functions are likely to fall in this category. Consider the Captricity example explored above. One might plausibly argue that it would be desirable to audit office-holders' financial data via a fully automated solution. But even so, because that ideal is not yet technologically possible, it becomes a matter of obvious public concern and accountability what types of shadow adjustments are taking place—at the behest of humans—behind the scenes.[76]

Temporary "bootstrapping" of human labor into not-yet-but-hopefully-someday-automated systems can also help us begin to understand how users are likely to interact with these systems.[77] This would allow for important research on human-computer interaction that can proceed alongside technical innovations. The "Wizard of Oz" experimental method, developed in the 1980s for human factors research, similarly involves a researcher controlling a system that a research subject believes to be autonomous, typically in order to study some aspect of the system that can be examined

76. *See* Griswold, *supra* note 49 and accompanying text (discussing the Capricity case in more detail).

77. Roboticist Wendy Ju uses this term to describe the Kiwibot's human support operation. Said, *supra* note 56.

without a fully built-out system.[78] Though researchers must always be attentive to the ethical implications of deception in research, such methods also permit much more rapid learning than would otherwise be possible.[79] But faux automation seemingly on its way to full automation can also be a fraudulent overpromise, as in the Theranos case.

2. The second possible dynamic is that, ideally, a decision-making system would neither be, nor appear to be, fully automated—but at present it has the veneer of automation.

Table 7: Unrealized Ideal—Manifest Humanity

	Human is in the loop	**Human is not in the loop**
Human appears to be in the loop	UNREALIZED IDEAL (manifest humanity)	
Human does not appear to be in the loop	PRESENT STATUS QUO (faux automation)	

This dynamic gives rise to a different set of governance questions. In essence, are there benefits associated with making actually non-automated systems look and feel more automated? We suspect the answer is almost always going to be no, for at least two reasons. The first is a simple anti-deception rationale; liberal subjects are entitled to know how the world they occupy actually works. Second, in decision-making environments that involve human judgment, we almost always care about which humans are entrusted to do the judging (and whom ought to be held to account for its outcomes). By necessity, a veneer of automation shuts that inquiry down.

Here, a good example may be content moderation. One could argue that First Amendment principles not only counsel in favor of continued human involvement in decisions about what content is so offensive or otherwise harmful that it merits restriction, but also compel us to reveal that human involvement to users. Doing so is the only way to surface the reality and dignity of the human labor required to support a system and to govern appropriately around it.

78. Paul Green & Lisa Wei-Haas, *The Rapid Development of User Interfaces: Experience With the Wizard of Oz Method*, 29 HUM. FACTORS SOC'Y, 470, 470–74 (1985).

79. Westlund & Breazeal, *supra* note 64; Hartzog, *supra* note 36, at 793–96.

3. The third possible dynamic is that, ideally, a decision-making system would involve human input, but appear to be fully automated, making the faux automation quadrant itself the optimum.

Table 8: Realized Ideal

	Human is in the loop	**Human is not in the loop**
Human appears to be in the loop		
Human does not appear to be in the loop	PRESENT STATUS QUO & REALIZED IDEAL (faux automation)	

We confess to having difficulty imagining cases that might actually populate this category and include it mostly for the sake of analytic symmetry. Nevertheless, it is possible that some cases do, or will, fall into this bucket.[80] For instance—and acknowledging the relatively far-flung nature of these examples—faux automation might be appropriate in situations where human input is desired, but where the source or nature of the input needs to be obscured. By analogy, one might think of firing squads as a kind of faux automation designed to obscure the source of human input: no one can tell which member of the squad is directly responsible for the fatality (and traditionally, one of the squad's rifles is loaded with blank cartridges to further permit each individual to disclaim moral responsibility).[81] This phenomenon is also exemplified in *per curiam* opinions, a judicial practice designed to achieve somewhat similar effects. In *per curiam* opinions, the opinion is considered to be rendered by *the court*, not by any specific judge. While these are crude approximations of cases that would actually call for the kinds of faux automation discussed in this paper, they give us reason to believe such cases—where the ideal involves disavowing but nonetheless maintaining a "human hand"—might exist. So for the moment, we leave the question open.

80. *See supra* at Part V.A.3.

81. Hanny Hindi, *Take My Life, Please*, SLATE (May 5, 2006), https://slate.com/news-and-politics/2006/05/merciful-but-messy-alternatives-to-lethal-injection.html [https://perma.cc/U3UN-T293].

VI. CONCLUSION

The age of automation is upon us. As more and more traditionally-human tasks become the province of machines, questions of governance loom large. These questions will be difficult enough in settings where the status of automation is apparent. But they will become even thornier in settings where the actuality and appearance of decision-making systems are misaligned.

In sketching our taxonomy of potential dynamics produced by misalignment, we mean to raise questions rather than resolve them. Put simply, the idea is that *any* time we are confronted with faux automation or skeuomorphic humanity, there will be at least two issues on the table. First, what kind of dynamic are we dealing with—in other words, what is the desirable end state? Second, how should we proceed within the context of that dynamic?

Both questions demand public deliberation and democratic oversight. This ideal is not always borne out in practice, for many reasons: it is costly; it relies on often-scarce political will; it becomes, at times, functionally impossible. Our point is that democratic oversight always matters in principle, even when it proves difficult in practice, and that misalignment is risky in large part because it stands to undermine such oversight. In the case of skeuomorphic humanity, the worry is that we—in the sense both of individual affected parties and of the public writ large—will be lulled, by a false sense of familiarity, into passively accepting inadvisable forms of automation. In the case of faux automation, by contrast, the worry is that we (again, in both senses) will be misled about automation's promise. We will not be able to coherently assess the costs and benefits of automation when its operation seems too good to be true.

The upshot is not that skeuomorphic humanity and faux automation are always lamentable. Each may have desirable features that override concerns about deception in particular situations. But weighing the harms of deception against other context-specific values requires knowing that deception is going on in the first place. Not only is misalignment poised to sow confusion and alienation, it's also liable, perversely, to thwart the very cost-benefit inquiry required to decide whether misalignment itself is permissible.

Going forward, the question of when misalignment is permissible—and if not, what constitutes the proper remedy—will be complex and unlikely to yield easy answers. This does not make the questions intractable. It simply requires public deliberation and democratic oversight. The future of automation, including the interplay between reality and appearance, must be something we resolve together through policy—not something imposed on us.