6

Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions

Daniel N. Kluttz, Nitin Kohli, and Deirdre K. Mulligan[*]

INTRODUCTION

Offering a "barously brief" distillation of Marshall McLuhan's writings, John M. Culkin expanded on one of McLuhan's five postulates, *Art Imitates Life*, with the now-famous line, *We shape our tools and thereafter they shape us*.[1] This fear of being shaped and controlled by tools, rather than autonomously wielding them, lies at the heart of current concerns with machine learning and artificial intelligence systems (ML/AI systems). Stories recounting the actual or potential bad outcomes of seemingly blind deference and overreliance on ML/AI systems crowd the popular press. Whether it is Facebook's algorithms allowing Russian operatives to unleash a weapon of mass manipulation, trained on troves of personal data, on electorates in the US and other countries; inequitable algorithmic bail decisions placing people of color behind bars while whites with similar profiles are sent home to await trial; cars in autonomous mode driving their inattentive could-be-drivers to their death; or algorithms assisting Volkswagen in routing around air quality regulations, there is a growing sense that our tools, if left unchecked, will undermine our choices, our values, and our public policies.

If we fail to grapple with the significant challenges posed by ML/AI systems designed to automate tasks or aid decision making, things may get much worse. At risk are potential decreases in human agency and skill,[2] both over- and under-reliance on decision support systems,[3] confusion about

[*] Titles in alphabetical order.

[1] Culkin, J. M. 1967. "A Schoolman's Guide to Marshall McLuhan." *The Saturday Review*, March 1967, 51–53, 70–72.

[2] Lee, John D., and Bobbie D. Seppelt. 2009. "Human Factors in Automation Design." In *Springer Handbook of Automation*, edited by Shimon Nof, pp. 417–36. Springer: Berlin (detailing how automation that fails to attend to how it redefines and restructures tasks, and the behavioral, cognitive, and emotional responses of operators to these changes, produce various kinds of failure, including those that arise from deskilling due to reliance on automation).

[3] Goddard, Kate, Abdul Roudsari, and Jeremy C. Wyatt. 2012. "Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators." *Journal of the American Medical Informatics*

responsibility,[4] and diminished accountability.[5] Relatedly, as technology reconfigures work practices, it also shifts power in ways that may misalign with liability frameworks, diminishing humans' agency and control but still leaving them to bear the blame for system failures.[6] Automation bias, power dynamics, belief in the objectivity and infallibility of data, and distrust of professional knowledge and diminished respect for expertise – all coupled with the growing availability of ML/AI systems and services – portend a potential future in which we are *ruled by our tools*.

Designing a future in which our tools help us reason and act more effectively, efficiently, and in ways aligned with our social values – i.e., creating the tools that help us act responsibly – requires attention to system design and governance models. ML/AI systems that support us, rather than control us, require designs that foster in-the-moment human engagement with the knowledge and actions systems produce, and governance models that support ongoing critical engagement with ML/AI processes and outputs. Expert decision-support systems are a useful case study to consider the system properties that could maintain human engagement and the governance choices that could ensure they emerge.

We begin by describing three new challenges – design by data, opacity to designer, and dynamic and variable features – posed by the use of predictive algorithmic systems in professional, expert domains. Concerns about inscrutable bureaucratic rules and privatization of public policy making (and the specific opacity that technology can bring to either) apply to predictive machine learning systems generally, but we suggest there are distinctive challenges posed by such predictive systems. We then briefly explore transparency and explainability, two policy objectives that current scholarship suggests are antidotes to such challenges. We show how conceptions of transparency and explainability differ along disciplinary lines (e.g., law, computer science, social

Association 19 (1): 121–27(reviewing literature on automation bias in health care clinical decision support systems); Bussone, A., S. Stumpf, and D. O'Sullivan. 2015. "The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems." In *2015 International Conference on Healthcare Informatics*, 160–69, p. 160 (discussing research findings on automation bias and self-reliance).

[4] For an overview of research on technology-assisted decision making and responsibility, see Mosier, Kathleen L., and Ute M. Fischer. 2010. "Judgment and Decision Making by Individuals and Teams: Issues, Models, and Applications." *Reviews of Human Factors and Ergonomics* 6 (1): 198–256.

[5] Nissenbaum, Helen. 1994. "Computing and Accountability." *Commun. ACM* 37 (1): 72–80; Simon, Judith. 2015. "Distributed Epistemic Responsibility in a Hyperconnected Era." In *The Onlife Manifesto: Being Human in a Hyperconnected Era*, edited by Luciano Floridi, pp. 145–59. Cham, CH: Springer International Publishing.

[6] Jones, Meg Leta. 2015. "The Ironies of Automation Law: Tying Policy Knots with Fair Automation Practices Principles." *Vanderbilt Journal of Entertainment & Technology Law* 18 (1): 77–134; Elish, Madeleine C. 2016. "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction." In *We Robot 2016 Working Paper*, 1–26. University of Miami (exploring how humans take the brunt of failures in sociotechnical systems, acting as "moral crumple zones" and absorbing a disproportionate amount of responsibility and liability, and arguing for reapportioning responsibility and liability in relation to actual control and agency).

sciences) and identify limitations of each concept for addressing the challenges posed by algorithmic systems in expert domains.

We then introduce the concept of contestability and explain the particular benefits of contestable ML/AI systems in the professional context over and above transparent or explainable systems. This approach can be valuable for an algorithmic handoff in a highly professionalized domain, such as the use of predictive coding software – a particular e-discovery tool – by lawyers during litigation. Current governance frameworks around the use of predictive coding in the form of professional norms and codified rules and regulations have their limitations. We argue that an approach centered around contestability would better promote attorneys' continued, active engagement with these algorithmic systems without relying so heavily on retrospective, case-specific, and costly legal remedies.

THE LIMITATIONS OF EXISTING APPROACHES TO PROTECTING VALUES

Technical systems containing algorithms are shaping and displacing human decision making in a variety of fields, such as criminal justice,[7] medicine,[8] product recommendations,[9] and the practice of law.[10] Such decision-making handoffs have been met with calls for greater transparency and explainability about system-level and algorithmic processes. The delegation of *professional* decision making to predictive algorithms – models that predict or estimate an output based on a given input[11] – creates additional issues with respect to opacity in machine learning[12] and to more general concerns with bureaucratic inscrutability[13] and privatization of public power.[14]

[7] Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias." *ProPublica*, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[8] See, e.g., Faggella, Daniel. 2018. "Machine Learning Healthcare Applications – 2018 and Beyond." *TechEmergence*. March 1, 2018. https://www.techemergence.com/machine-learning-healthcare-applications/; see generally, Berner, Eta S., ed. 2016. *Clinical Decision Support Systems: Theory and Practice*. 3rd ed. Health Informatics. New York: Springer.

[9] As an example, see Netflix's recommendation engine: https://medium.com/netflix-techblog/netflix-recommendations-beyond-the-5-stars-part-1-55838468f429.

[10] Ashley, Kevin D. 2017. Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age. Cambridge: Cambridge University Press.

[11] James, G., D.Witten, T. Hastie, and R. Tibshirani. (2013). *An Introduction to Statistical Learning*. New York: Springer.

[12] Burrell, Jenna. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3 (1): 1–13 (describing three forms of opacity: corporate or state secrecy; technical illiteracy; and complexity and scale of machine-learning algorithms).

[13] Freeman, Jody. 2000. "Private Parties, Public Functions and the New Administrative Law Annual Regulation of Business Focus: Privatization." *Administrative Law Review* 52: 813–58.

[14] Citron, Danielle Keats. 2008. "Technological Due Process." *Washington University Law Review* 85 (6): 1249–313. ("Agencies inadvertently give rulemaking power to computer programmers who can, and do, alter established policy when embedding it into code." "Because the policies embedded in

Three Challenges Facing Algorithmic Systems in Expert Domains

We identify three challenges facing the use of predictive algorithms in expert systems. First, such predictive algorithms are not designed by technologists in the traditional sense. Whereas engineers of traditional expert systems explicitly program in a set of rules, ideally from the domain knowledge of adept individuals, predictive algorithms supplant this expert wisdom by deriving a set of decision rules from data.

Predictive algorithms can be partitioned into two categories: (1) those focused on outcomes that do not rely too heavily on professional judgment (e.g., was an individual readmitted to the hospital within thirty days of their visit?) versus (2) those focused on outcomes that are more tailored toward emulating the decisions made by professionals with specific domain expertise (e.g., does this patient have pneumonia?). Specifically, the first example can be deemed either true or false simply via observation of admit logs, regardless of professional training. The second example, by way of contrast, is distinct from the first in that it requires medical expertise to make such a diagnosis. In the strictest sense, expert systems fall into the second category,[15] and as such, inferences of such rules via predictive algorithms create unique challenges for the transfer of expertise from both individuals to the algorithm, and from the algorithm to individuals.

The second challenge is one of opacity. In many ways, this issue is induced by the first. While certain classes of predictive algorithms lend themselves to ease of understanding (such as logistic regression and shallow decision trees), other classes of model make it difficult to understand the rules inferred from the data (such as neural networks and ensemble methods). Unlike expert systems, where domain professionals can review and interrogate the internal rules, the opacity of certain algorithms prevents explicit examination of these decision rules, leaving experts to infer the model's underlying reasoning from input–output relationships.

Last, these algorithms are case-specific and evolving. They will not necessarily make the same decision about two distinct people in the same way at the same point in time, neither will they necessarily make the same decision about the same individual at varying points in time. This plasticity creates challenges for understanding and interrogating a model's behavior, as input–output behavior can vary from case to case and can vary over time.

Transparency: Perspectives and Limitations

Due to the challenges described above, algorithmic handoffs have been met with calls for greater transparency.[16] At a fundamental level, transparency refers to some

code are invisible, administrators cannot detect when the rules in an automated system depart from formal policy.")

[15] See Todd, Bryan S. 1992. *An Introduction to Expert Systems*. Oxford: Oxford University Computing Laboratory.

[16] Brauneis, Robert, and Ellen P. Goodman. 2018. "Algorithmic Transparency for the Smart City." *Yale Journal of Law & Technology* 20: 103–76, p. 108.

notion of openness or access, with the goal of becoming informed about the system. However, the word "transparency" lends itself to the question: *What* is being made transparent?

Given the growing role that algorithmically driven systems are poised to play across government and the private sector, we should exercise care in choosing policy objectives for transparency. A trio of federal laws – two adopted in the 1970s due to fears that the federal government was amassing data about citizens – exemplify three policy approaches to transparency relevant to algorithmic systems. Together, the laws aim to ensure citizens "know what their Government is up to,"[17] that "all federal data banks be fully and accurately reported to the Congress and the American people,"[18] that individuals have access to information about themselves held in such data banks, and that privacy considerations inform the adoption of new technologies that manage personal information. These approaches can be summarized as relating to (1) scope of a system, (2) the decision rules of a process, and (3) the outputs.

The Privacy Act of 1974,[19] which requires notices to be published in the Federal Register prior to the creation of a new federal record-keeping system, and section 208 of the E-Government Act of 2002,[20] which requires the completion of privacy impact assessments, exemplify the scope perspective. These laws provide notice about the existence and purpose of data-collection systems and the technology that supports them. For example, the Privacy Act of 1974 requires public notice that a system is being created and additional information about the system, including its name and location, the categories of individual and record maintained in the system, the use and purpose of records in the system, agency procedures regarding storage, retrieval, and disposal of the records, etc.[21] The first tenet of the Code of Fair Information Practices, first set out in a 1973 HEW (Health, Education, Welfare) Report[22] and represented in the Privacy Act of 1974 and data-protection laws the world over, stipulates in part that "there must be no personal-data record-keeping systems whose very existence is secret."[23] With the Privacy Act of 1974, the transparency theory is one of public notice and scope. Returning to our previous question of "what is being

[17] US Dept. of Justice v. Reporters Committee, 489 U.S. 749, 773 (1989).

[18] Ware, W. H., 1973. Records, Computers and the Rights of Citizens (No. P-5077). Santa Monica, CA: RAND Corporation.

[19] 5 U.S.C. § 552a (2014).

[20] Pub. L. No. 107–347, § 208, 116 Stat. 2899 (Dec. 17, 2002).

[21] 5 U.S.C. § 552a(e)(4); *see also* United States Department of Health, Education, and Welfare. 1973. "Report of the Secretary's Advisory Committee on Automated Personal Data Systems, Records, Computers, and the Rights of Citizens." MIT Press (discussing purpose and provisions of Privacy Act).

[22] US Department of Health, Education, and Welfare. "Report of the Secretary's Advisory Committee on Automated Personal Data Systems: Records, Computers and the Rights of Citizens," 1973, at § III. Safeguards for Privacy.

[23] The full Code of Fair Information Practices can be found at https://epic.org/privacy/consumer/code_fair_info.html.

made transparent," in this approach to transparency, it is precisely the existence and scope being made available.

Unlike the scope aspect of transparency, the decision-rules aspect is not concerned with whether or not such a system exists. Rather, this view of transparency refers to tools to extract information about how these systems function. As an example, consider the Freedom of Information Act (FOIA), a law that grants individuals the ability to access information and documents controlled by the federal government.[24] The transparency theory here is that the public has a vested interest in accessing such information. But instead of disclosing the information upfront, it sets up a mechanism to meet the public's demand for it. As such, FOIA allows for individuals to gain access to the decisional rules of these systems and processes. Similarly, the privacy impact assessment requirement of the E-Government Act of 2002 provides transparency around agencies' consideration of new technologies, as well as their ultimate design choices.

Last, several privacy laws allow individuals to examine the inputs and outputs of systems that make decisions about them. Under this perspective, transparency is not the end goal itself. Rather, transparency supports the twin goals of ensuring fair inputs and understanding the rationale for the outputs by way of pertinent information about the inputs and reasoning. The laws all entitle individuals to access information used about them and to correct or amend data. Some of the privacy laws in this area also entitle individuals to receive information about the reasons behind negative outcomes.[25] For example, under the Equal Credit Opportunity Act, if a candidate's credit application is rejected, the credit bureau must provide the key reasons for the decision.[26] Thus, this type of transparency refers to notice of how a particular decision was reached. These forms of transparency are aimed at individual, rather than collective, understanding; they provide, to a limited extent, insight into the data and the reasoning – or functioning – of systems.

Within the computer science literature, transparency is similar to the functional and outputs perspective presented in law. That is, transparency often refers to some notion of openness around either the internals of a model or system, or around the outputs. Typically, less focus is given to disclosing the subjective choices that were invoked during the system design and engineering process or to system inputs.

The social sciences and statistics, however, take a more comprehensive perspective on transparency. Transparency in these disciplines not only captures the ideas from law and computer science, but also means disclosures about how the data was gathered, how it was cleaned and normalized, the methods used in the analysis, the choice of hyperparameters and other thresholds, etc., often in line with the goals of

[24] 5 U.S.C. § 553 (2016).

[25] See, e.g., The Equal Credit Opportunity Act (ECOA), 15 U.S.C. § 1691 et seq., as implemented by Regulation B, 12 C.F.R. §1002.9. *See also* The Fair Credit Reporting Act (FCRA), 15 U.S.C. § 1681 et seq.

[26] 15 U.S.C. § 1691(d).

reproducibility.[27] The sweep of transparency reflects an understanding that these choices contribute to the methodological design and analysis. This more holistic approach to transparency acknowledges the effect that humans have in this process (reflected in decisions about data, as well as behaviors captured in the data), which is particularly pertinent for predictive algorithms.

Current policy debates, and scientific research, center around explainability and interpretability. Transparency is being reframed, particularly in the computer science research agenda, as an instrumental rather than final objective of regulation and system design. The goal is not to lay bare the workings of the machine, but rather to ensure that users understand how the machines are making decisions – whether those decisions be offering predictions to inform human action or acting independently. This reflects both growing recognition of the inability of humans to understand how some algorithms work even with full access to code and data, but also an emphasis on the overall system – rather than solely the *algorithm* – as the artifact to be known.

Explainability: Perspectives and Limitations

Explainability is an additional design goal for machine-learning systems. Driven in part by growing recognition of the limits of transparency to foster human understanding of algorithmic systems, and in part by pursuit of other goals such as safety and human compatibility, researchers and regulators are shifting their focus to techniques and incentives to produce machine-learning systems that can explain themselves to their human users. Such desires are well-founded in the abstract. For the purposes of decision making or collaboration, explanations can act as an interface between an end-user and the computer system, with the purpose of keeping a human in the loop for safety and discretion. Hence, explanations invite questioning of AI models and systems to understand limits, build trust, and prevent harm. As with transparency, different disciplines have responded to this call to action by operationalizing both explanations and explainability in differing ways.

One notable use of explanations and explainability comes from the social sciences. Miller[28] performed a comprehensive literature review of over 200 articles from the social sciences and found that explanations are causal, contrastive, selective, and social. What is pertinent from this categorization is how well the paradigms invoked in predictive algorithms (machine learning, artificial intelligence, etc.) fall within social understandings of explanations. Machine learning raises difficulties for all four of Miller's attributes of explanations.

[27] Miguel, Edward, Colin Camerer, Katherine Casey, Joshua, Cohen, Kevin M. Esterling, Alan Gerber, Rachel Glennerster, et al. 2014. "Promoting Transparency in Social Science Research." *Science* 343 (6166): 30–31.

[28] Miller, Tim. 2017. "Explanation in Artificial Intelligence: Insights from the Social Sciences." *ArXiv:1706.07269 [Cs]*, June. http://arxiv.org/abs/1706.07269.

For concreteness and clarity, imagine we have a predictive algorithm that classifies a patient's risk for breast cancer as either low risk, medium risk, or high risk. In this scenario, a causal explanation would answer the question: "Why was the patient classified as high risk?" Alternatively, a contrastive explanation would answer questions of the form, "Why was the patient classified as high risk as opposed to low risk or medium risk?" As such, explanations of the causal type require singular scope on the outcome, whereas contrastive explanations examine not only the predicted outcome, but other candidate alternatives as well.

With respect to machine learning, this distinction is important and suggestive. Machine learning is itself a correlation box. As such, the *output itself should not be interpreted as causal*. However, when individuals ask for *causal explanations* of predictive algorithms, they are not necessarily assuming that the underlying data mechanism is causal. Rather, the notion of causality is seeking to understand what caused the algorithm to decide that the patient was high risk, not what caused the patient to be high risk in actuality. Thus, causal explanations can be given of a model built on correlation. However, the fact that they can be produced doesn't mean that causal explanations further meaningful understanding of the system.

Contrastive explanations are a better fit for machine learning. The very paradigm of machine learning – classification models – are built in a contrastive manner. These models are trained to learn to pick the "best" output given a set of inputs – or equivalently stated, the model is taught to discern an answer to a series of input questions based on the fixed set of alternatives available. Combining these insights, it follows that requiring causal explanations for classification models is inappropriate for determining why a model predicted the value it did. Contrastive explanations, which provide insight into the counterfactual alternatives that the model rejected as viable, transfer more knowledge about the system, than causal ones.

Regardless of whether the type of explanation is causal or contrastive, Miller argued that explanations in the social sciences were selective. That is, explanations tend to highlight a few key justifications rather than being completely exhaustive. Consider the case of a doctor performing a breast cancer-screening test in the absence of a predictive algorithm. When relaying the rationale of their diagnosis to a patient, a doctor would provide sufficient reasons for their decision to justify their answer. Now, consider the state of the world where a handoff has been made to the predictive model. Suppose the model being used relies on 500 features. When explaining why the model predicted the outcome it did, it is indeed unreasonable to assume that providing information about all 500 features would practically relay any information about why the model made the choice it did. As such, requiring explanations of predictive models requires honing into the relevant features of a decision problem, which may differ from patient to patient and may vary over time.

On the aspect of explanations being social, Miller noted that explanations are meant to transfer knowledge from one individual to another. In the example above, where the doctor performs the breast cancer-screening test, this was the point of

having the doctor justify their diagnosis to the patients – to inform the patient about their breast cancer-risk level. When applied to technical systems, the goal is to transfer knowledge about the internal logic of how the system reached its conclusion to some individual (or class of individuals). In the case of our breast cancer-risk prediction, this would manifest itself as a way to justify why the algorithm predicted high risk as opposed to low risk. It is worth noting that for predictive algorithms, it is often difficult to truly achieve the social goal of explanations. Certain qualities of algorithms – such as their functional form (e.g., nonlinear, containing interaction terms), their input data, and other characteristics – make it particularly difficult to assess the internal logic of the algorithm itself, or for the system to even explain what it is doing. It is therefore difficult for these machine systems to transfer knowledge to individuals in the form of an explanation that is either causal or contrastive. To the extent that explanations are aimed at improving human understanding of the logic of algorithms, the qualities of some algorithms may be incompatible with this means of transferring knowledge. It may be that the knowledge transfer must come the other way around, from the human to the machine, which is then bound to particular way or ways of knowing.[29]

Thus, there are tensions between the paradigms of predictive algorithms and those characteristics laid out by Miller. As such, the discussion above suggests that our target is off. That is, to actually fully and critically engage with predictive algorithms, this suggests that we require something stronger than transparency and explainability. Enter *contestability* – the ability to challenge machine predictions.

TOWARD CONTESTABILITY AS A FEATURE OF EXPERT DECISION-SUPPORT SYSTEMS

Contestability fosters engagement rather than passivity, questioning rather than acquiescence. As such, contestability is a particularly important system quality where the goal is for predictive algorithms to enhance and support human reasoning, such as decision-support systems. Contestability is one way "to enable responsibility in knowing"[30] as the production of knowledge is spread across humans and machines. Contestability can support critical, generative, and responsible engagement between users and algorithms, users and system designers, and ideally between users and those subject to decisions (when they are not the users), as well as the public.

[29] Kroll, Joshua A., Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. 2017. "Accountable Algorithms." *University of Pennsylvania Law Review* 165 (3): 633–705.

[30] Simon, Judith. 2015. "Distributed Epistemic Responsibility in a Hyperconnected Era." *The Onlife Manifesto*, pp. 145–59. Cham, CH: Springer International Publishing, at p. 146 (separating out two aspects of "epistemic responsibility": 1) the individualistic perspective, which asks, "what does it mean to be responsible in knowing?"; and 2) the governance perspective with asks, "what does it take to enable responsibility in knowing?").

Efforts to make algorithmic systems knowable respond to the individual need to understand the tools one uses, as well as the social need to ensure that new tools are fit for purpose. Contestability is a design intervention that can contribute to both.[31] However, our focus here is on its potential contribution to the creation of governance models that "support epistemically responsible behavior"[32] and support shared reasoning about the appropriateness of algorithmic systems behavior.[33]

Contestability, the ability to contest decisions, is at the heart of legal rights that afford individuals access to personal data and insight into the decision-making processes used to classify them,[34] and it is one of the interests that transparency

[31] For insights on how contestable systems advance individual understanding, see, e.g., Eslami, Motahhare, and Karrie Karahalios. 2017. "Understanding and Designing around Users' Interaction with Hidden Algorithms in Sociotechnical Systems." *CSCW Companion* (describing several studies finding that seamful designs, which expose algorithmic reasoning to users, facilitated understanding, improved user engagement, and in some instances altered user behavior); Eslami, Motahhare, et al. 2015. "I Always Assumed that I Wasn't Really That Close to [Her]: Reasoning about Invisible Algorithms in News Feeds." *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (describing the lasting effects on how users engage with Facebook to influence the News Feed algorithm after an experimental design intervention that visualized its curatorial voice); Jung, Malte F., David Sirkin, and Martin Steinert. 2015. "Displayed Uncertainty Improves Driving Experience and Behavior: The Case of Range Anxiety in an Electric Car." *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15) (gradient plot that reveals uncertainty reduced anxiety over single point estimate of remaining range of electric vehicle); Joslyn, Susan, and Jared LeClerc. 2013. "Decisions with Uncertainty: The Glass Half Full." *Current Directions in Psychological Science* 22 (4): 308–15 (displaying uncertainty in weather predictions can lead to more optimal decision making and trust in a forecast: transparency about probabilistic nature of prediction engenders trust even when predictions are wrong); Stumpf, Simone, et al. 2007. "Toward Harnessing User Feedback for Machine Learning." *Proceedings of the 12th International Conference on Intelligent User Interfaces*; Stumpf, Simone, et al. 2009. "Interacting Meaningfully with Machine-Learning Systems: Three Experiments." *International Journal of Human-Computer Studies* 67 (8): 639–62 (explainable systems can improve user understanding and use of system and enable users to provide deep and useful feedback to improve algorithms); Moor, Travis, et al. 2009. "End-User Debugging of Machine-Learned Programs: Toward Principles for Baring the Logic" (salient explanations helped users adjust their mental models); Amershi, Saleema, et al. 2014. "Power to the People: The Role of Humans in Interactive Machine Learning." *AI Magazine* 35 (4): 105–20 (providing an overview of interactive machine learning research, with case studies, and discussing value of interactive machine learning approaches for machine learning community as well as users).

[32] Simon, Judith. 2015. "Distributed Epistemic Responsibility in a Hyperconnected Era." In *The Onlife Manifesto: Being Human in a Hyperconnected Era*, edited by Luciano Floridi, pp. 145–59. Cham, CH: Springer International Publishing, at p. 158.

[33] Reuben Binns argues that "*public reason*—roughly, the idea that rules, institutions and decisions need to be justifiable by common principles, rather than hinging on controversial propositions which citizens might reasonably reject—is an answer to the problem of reasonable pluralism in the context of algorithmic decision making," and requires transparency. Binns, Reuben. 2017. "Algorithmic Accountability and Public Reason." *Philosophy & Technology*, May.

[34] See, e.g., regulations under the notification provisions of the Equal Credit Opportunity Act 15 U.S.C. § 1691 et seq. that require those denied credit to be provided specific, principal reasons for the denial ECOA 12 C.F.R. § 1002.1, et seq. at §1002.9; Hildebrandt, M. 2016. "The New Imbroglio. Living with Machine Algorithms." In *The Art of Ethics in the Information Society*, edited by L. Janssens, 55–60. Amsterdam: Amsterdam University Press, p. 59 (arguing that the EU General Data Protection

serves. Contestability as a design goal, however, is more ambitious and far-reaching. A system designed for contestability would protect the ability to contest a specific outcome, consistent with privacy and consumer protection law. It would also facilitate generative engagement between humans and algorithms throughout the use of the machine-learning system and support the interests and rights of a broader range of stakeholders – users, designers, as well as decision subjects – in shaping its performance.

Hirsch et al. set out contestability as a design objective to address myriad ethical risks posed by the potential reworking of relationships and redistribution of power caused by the introduction of machine-learning systems.[35] Based on their experience designing a machine-learning system for psychotherapy, Hirsch et al. offer three lower-level design principles to support contestability: (1) improving accuracy through phased and iterative deployment with expert users in environments that encourage feedback; (2) heightening legibility through mechanisms that "unpack aggregate measures" and "trac[e] system predictions all the way down" so that "users can follow, and if necessary, contest the reasoning behind each prediction"; and relatedly, in an effort to identify and vigilantly prevent system misuse and implicit bias, (3) identifying "aggregate effects" that may imperil vulnerable users through mechanisms that allow "users to ask questions and record disagreements with system behavior" and engage the system in self-monitoring.[36] Together, these design principles can drive active, critical, real-time engagement with the reasoning of machine-learning system inputs, outputs, and models.

This sort of deep engagement and ongoing challenge and recalibration of the reasoning of algorithms is essential to yield the benefits of humans and machines reasoning together. Concerns that engineers will stealthily usurp or undermine the decision-making logics and processes of other domains have been an ongoing and legitimate complaint about decision support and other computer systems.[37]

Regulation requires "[Algorithmic] decisions that seriously affect individuals' capabilities must be constructed in ways that are comprehensible as well as contestable. If that is not possible, or, as long as this is not possible, such decisions are unlawful.") However, in reality, what the GDPR requires may be much more limited. *See also* Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. 2017. "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation." *International Data Privacy Law* 7 (2): 76–99, p. 93 (arguing that a fairer reading of the GDPR provisions and recitals, and member states implementation of the EU Data Protection Directive it replaces, would require "limited disclosures of the 'logic involved' in automated decision making, primarily concerning system functionality rather than the rationale and circumstances of specific decisions").

35 Hirsch, Tad, Kritzia Merced, Shrikanth Narayanan, Zac E. Imel, and David C. Atkins. 2017. "Designing Contestability: Interaction Design, Machine Learning, and Mental Health." *DIS. Designing Interactive Systems (Conference)* 2017 (June): 95–99 (describing the way an automated assessment and training tool for psychotherapists could be used as a "blunt assessment tool" of management to the detriment of therapists and patients) at p. 98.

36 Id. at p. 98.

37 See Citron, Danielle Keats. 2008. "Technological Due Process." *Washington University Law Review* 85 (6): 1249–313 (identifying the slippage and displacement of case worker values by engineering rules embedded in an expert system); Moor, James H. 1985. "What Is Computer Ethics?" *Metaphilosophy*

Encouraging human users to engage and reflect on algorithmic processes can reduce the risk of stealthy displacement of professional and organizational logics by the logics of software developers and their employers. Where an approach based on explanations imagines questioning and challenging as out-of-band activities – exception handling, appeals processes, etc. – contestable systems are designed to foster critical engagement within the system. Such systems use that engagement to iteratively identify and embed domain knowledge and contextual values, as decision making becomes a collaborative effort within a sociotechnical system.

In the context of decision-support systems, increasing system explainability and interpretability is viewed as a strategy to address errors that stem from automation bias and to improve trust.[38] Researchers have examined the impact of various forms of explanatory material, including confidence scores, and comprehensive and selective lists of important inputs, on the accuracy of decisions, deviation from system recommendations, and trust.[39] The relationship between explanations and correct decision making is not conclusive.[40]

Policy debates, like the majority of research on interpretable systems, envision explanations as static.[41] Yet, the responsive and dynamic tailoring at which machine learning and AI systems excel could allow explanations to respond to the expertise and other context-specific needs of the user, yielding decisions that leverage, and iteratively learn from, the situated knowledge and professional expertise of users.

16 (4): 266–75 (identifying three ways invisible values manifest in technical systems – to hide immoral behavior, gap-filling during engineering that invisibly embeds coders' value choices, and through complex calculations that defy values analysis); Burrell, Jenna. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3 (1): 1–13 (describing three forms of opacity in corporate or state secrecy, technical illiteracy, and complexity and scale of machine-learning algorithms).

38 Nunes, Ingrid, and Dietmar Jannach. 2017. "A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems." *User Modeling and User-Adapted Interaction* 27 (3–5): 393–444 (reviewing approaches to explanations in "advice-giving systems"); Bussone, A., S. Stumpf, and D. O'Sullivan. 2015. "The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems." In *2015 International Conference on Healthcare Informatics*, 160–69.

39 Bussone et al. 2015, supra note 38.

40 Id. at 161 (describing different research finding explanations leading to better and worse decisions).

41 Abdul, Ashraf, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. "Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda." In *Proceedings of the International Conference on Human Factors in Computing Systems*, 1–18 CHI '18 (research review concluding that the explainable AI research community generally produces static explanations focused on conveying a single message and recommending that research explore interactive explanations that allow users to more dynamically explore and interact with algorithmic decision-making systems); but see also Nunes, Ingrid, and Dietmar Jannach. 2017. "A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems." *User Modeling and User-Adapted Interaction* 27 (3–5): 393–444, p. 408 (describing research on interactive explanations that engage users by providing a starting point and allow them to probe systems through "(i) *what-if* (what the output would be if alternative input data were provided); (ii) *why* (why the system is asking for a particular input); and (iii) *why-not* (why the system has not provided a given output)" approaches).

The human engagement contestable systems invite would align well with regulatory and liability rules that seek to keep humans in the loop. For example, the Food and Drug Administration is directed to exclude from the definition of "device" those clinical decision support systems whose software function is intended for the purpose of:

> supporting or providing recommendations to a health care professional about prevention, diagnosis, or treatment of a disease or condition; and enabling [providers] to independently review the basis for such recommendations . . . so that it is not the intent that such [provider] rely primarily on any of such recommendations to make a clinical diagnosis or treatment decision regarding an individual patient.[42]

By excluding systems that prioritize human discretion from onerous medical-device approval processes, Congress shows its preference for human expert reasoning. Similarly, where courts have found professionals exhibiting overreliance on tools, they have structured liability to foster professional engagement and responsibility.[43] Systems designed for contestability invite engagement rather than delegation of responsibility. They can do so through both the provision of different kinds of information and an interactive design that encourages exploration and querying.

Professionals appropriate technologies differently, employing them in everyday work practice, as informed by routines, habits, norms, values and ideas and obligations of professional identity. Drawing attention to the structures that shape the adoption of technological systems opens up new opportunities for intervention. Appropriate handoffs to, and collaborations with, decision-support systems demand that they reflect professional logics and provide users with the ability to understand, contest, and oversee decision making. Professionals are a potential source of governance for such systems, and policy should seek to exploit and empower them, as they are well-positioned to ensure ongoing attention to values in handoffs and collaborations with machine-learning systems.

Regulatory approaches should seek to put professionals and decision support systems in conversation, not position professionals as passive recipients of system wisdom who must rely on out-of-system mechanisms to challenge them. For these reasons, calls for explainability fall short and should be replaced by regulatory approaches that drive contestable design. This requires attention to both the *information* demands of professionals – what they need to know such as training data, inputs, decisional rules, etc. – and *processes* of interaction that elicit professional expertise and allow professionals to learn about and shape machine decision making.

[42] 21 U.S.C. § 360j(o)(1)(E)(ii)-(iii) (2016); the term "device" is defined in 21 U.S.C. § 321(h).

[43] *Aetna Cas. and Sur. Co. v. Jeppesen & Co.*, 642 F.2d 339, 343 (9th Cir. 1981) (rejecting district court finding that pilots who relied on map that was defectively designed (showing topographical and elevation in distinct scales) were not negligent, because it would endorse a standard of care that would consider "pilot reliance on the graphics of the chart and complete disregard of the words and figures accompanying them" "as reasonable attention to duty by a pilot of a passenger plane" and opting instead to apportion fault).

Contestable Design Directions

Contestable design is a research agenda, not a suite of settled techniques to deploy. The question of what information and interactions will prompt appropriate engagement and shaping of a predictive coding system by professionals is likely to be both domain- and context-specific. However, there are systems in use and under development that support real-time questioning, curiosity, and scrutiny of machine learning systems' reasoning. First, Google's People and AI Research (PAIR) Initiative's "What-if Tool" is an actual tool that allows users to explore a machine-learning model. For example, users can see how changes in aspects of a dataset influence the learned model, understand how different models perform on the same dataset, compare counterfactuals, and test particular operational constraints related to fairness.[44] Second, LIME (Local Interpretable Model-agnostic Explanations), which generates locally interpretable models to explain the outputs of predictive systems, and SP-LIME, which builds on LIME to provide insight into the model (rather than a given prediction) by identifying and explaining a set of representative instances of the model's performance, offer information that, if presented to users, could inform their interaction with the model.[45] While the tools themselves focus only on surfacing information about decisions and models, if integrated with an interactive user interface, they could promote the explorations of predictions and models necessary for sound use of predictive systems to inform professional judgement.

Other research is exploring the ways in which structured interaction between domain experts and predictive models can improve performance.[46] There are two distinct approaches. One approach enables interaction during the development process. Here, the machine-learning training process is reframed as an HCI task, allowing a set of users the ability to iteratively refine a model during its conception.[47] In contrast to interaction during the development process, the second approach has focused on ways in which subject matter experts, with domain-specific knowledge, can interact with predictive systems that have already been developed in real time to invoke collaboration, exploration of data, and introspection.[48] At the very least,

[44] Wexler, James. 2018. "The What-If Tool: Code-Free Probing of Machine Learning Models." *Google AI Blog* (blog). September 11, 2018. http://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html. For the tool's code repository, see https://pair-code.github.io/what-if-tool/.

[45] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–144. KDD '16. New York: ACM.

[46] Micallef, Luana, Iiris Sundin, Pekka Marttinen, Muhammad Ammad-ud-din, Tomi Peltola, Marta Soare, Giulio Jacucci, and Samuel Kaski. 2017. "Interactive Elicitation of Knowledge on Feature Relevance Improves Predictions in Small Data Sets." In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, 547–52. IUI '17. New York: ACM.

[47] Dudley, John J. and Per Ola Kristensson. 2018. "A Review of User Interface Design for Interactive Machine Learning." *ACM Transactions on Interactive Intelligent Systems* 8 (2): 8:1–8, 37.

[48] Chander, Ajay, Ramya Srinivasan, Suhas Chelian, Jun Wang, and Kanji Uchino. "Working with Beliefs: AI Transparency in the Enterprise." In *Explainable Smart Systems Workshop, Intelligent User Interfaces*. 2018.

ensuring that decisions about things such as thresholds are decided by professionals in the context of use (and remain visible to those using the system), rather than set as defaults, can support greater engagement with predictive systems.

CONCLUSION

Contestability allows professionals, not just data, to train systems. In doing so, contestability transfers knowledge about how the machine is reasoning to the professional, and it allows the professional to collaborate, critique, and correct the predictive algorithm. While relevant professional norms, ethical obligations, and laws are necessary, design has a role to play in promoting responsible introduction of predictive ML/AI systems in professional, expert domains. Such systems must be designed with contestability in mind from the outset. Designing for contestability has some specific advantages compared to rules and laws. Opportunities to reflect on the inputs and assumptions that shape systems can avert disasters where they misalign with the conditions or understandings of professional users. Reminders of professional responsibilities and potential risks of not complying with them can prompt engagement before undesirable outcomes occur. Contestable design can confer training benefits allowing users to learn through use. Finally, it can be used to signal the distribution of responsibility from the start rather than relying solely on litigation to retrospectively mete it out in light of failures. Contestability can foster professional engagement with tools rather than deferential reliance. To the extent the goal is to yield the best of human-machine knowledge production, designing for contestability can promote the responsible production of knowledge with machine learning tools within professional contexts.