

ARTICLE

Do We Need to Know What Is Artificial? Unpacking Disclosure & Generating Trust in an Era of Algorithmic Action

Orly Lobel

Should users have the right to know when they are chatting with a bot? Should companies providing generative AI applications be obliged to mark the generated products as AI-generated or alert users of generative chats that the responder is “merely an LLM (or a Large Language Model)”? Should citizens or consumers—patients, job applicants, tenants, students—have the right to know when a decision affecting them was made by an automated system? Should art lovers, or online browsers, have the right to know that they are viewing an AI-generated image?

As automation accelerates and AI is deployed in every sector, the question about knowing about artificiality becomes relevant in all aspects of our lives. This essay, written for the 2024 Network Law Review Symposium on the Dynamic of Generative AI, aims to unpack the question—which is in fact a set of complex questions—and to provide a richer context and analysis than the often default, absolute answer: *yes/we, the public, must always have the right to know what is artificial and what is not*. The question is more complicated and layered than may initially seem. The answer, in turn, is not as easy as some of the recent regulatory initiatives suggest in their resolute yes. The answer, instead, depends on the goals of information disclosure. Is disclosure a deontological or dignitarian good, and in turn, right, in and of itself? Or does disclosure serve a utilitarian purpose of supporting the goals of the human-machine interaction, for example, ensuring accuracy, safety, or unbiased decision-making? Does disclosure increase trust in the system, process, and results? Or does the disclosure under certain circumstances hinder those very goals, for example, if knowing that a decision was made by a bot reduces the AI user’s trust and increases the likelihood the AI user will disregard the recommendation (e.g., an AI radiology or insulin bolus system recommendation? An AI landing device in aviation?).

The essay presents a range of contexts and regulatory requirements centered around the right to know about AI involvement. It then suggests a set of reasons for disclosure of artificiality: dignity; control; trust – including accuracy, consistency, safety, fairness; authenticity; ownership/attribution, and aesthetic/experiential. The essay further presents recent behavioral literature on AI rationality, algorithmic aversion, and algorithmic adoration to suggest a more robust framework within which the question about disclosure rights, and their effective timing, should be answered. It then shows how labeling and marking AI-generated images is a distinct inquiry separate from disclosure of AI-generated decisions. In each of these contexts, the answers should be based on empirical evidence on how disclosures affect perception, rationality, behavior, and measurable goals of these deployed technologies.

1. Awareness v. Trust

The lines between human and machine are blurring. The idea of neatly separating what is natural or human-made and that which is artificial and machine-made has been a longstanding collective fantasy in human thought. But today, digital technology, computational abilities, big data, and LLMs have accelerated the processes of human-machine symbiosis. AI assistants whisper recipes in our kitchens, chatbots answer our customer service woes, and algorithmically curated news feeds shape our understanding of the world. AI’s potential for increasing productivity, convenience, and efficiency has pushed developers to implement it in various aspects of everyday life. AI is used for customer service roles to address questions, complaints, and other inquiries customers may have. Beyond chatbots, AI is used in a range of ways in every sector and field of market and in social interactions to generate content, images, texts, and videos and to make predictions and decisions. In [education](#), AI is being developed to support teacher-student interaction through educational games, learning platforms, grading and feedback systems, student support chatbots, and tutoring.¹ In [healthcare](#), AI is used to alleviate administrative burdens, assist nurses, reduce dosage errors, compile health data quickly, and even help prevent insurance fraud.² AI is also now deployed in [e-](#)

¹ Ilana Hamilton & Brenna Swanston, *Artificial Intelligence in Education: Teachers’ Opinions On AI In The Classroom*, Forbes (Dec. 5, 2023), <https://www.forbes.com/advisor/education/artificial-intelligence-in-school/>.

² IBM Education, *The benefits of AI in healthcare*, IBM (July 11, 2023), <https://www.ibm.com/blog/the-benefits-of-ai-in-healthcare/>.

commerce,³ gaming,⁴ astronomy,⁵ financial services,⁶ cybersecurity,⁷ social media,⁸ agriculture,⁹ and in a range of government services.¹⁰ In this new reality, a key question arises: do we need to know if we're interacting with a machine?

Indeed, since AI can be deployed in a link of the value chain and may not necessarily be consumer-facing, the idea that we are living in a sharp binary world of choice between human and artificial is often too simplistic. If one considers that AI can be one of many "upstream inputs" of a consumer-facing service/product/output, how do we even draw the line between human and artificial? Should we know when we interact not with a machine, but with a human who is taking 100%- or about 90% or 50%... of a machine generated output? Is knowledge of the distinct functions and roles of these co-players necessary?

With the rise of AI there is also a global race to regulate it, with the aim of mitigating potential risks and harms. One aspect of regulation is whether public policy should mandate, or at least direct companies to disclose to users the artificiality of AI outputs. Does knowing whether a chatbot or decision-maker is artificial help protect against and mitigate risks? The right to know that you are interacting with a bot, or that you are subject to automated decision-making, is a centerpiece of EU/US legislative proposals. Both the EU's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) already include the right to know if AI is making decisions about them and to request explanations for those decisions. Under the new EU AI Act, consumers will have a right to see disclosures that they are chatting with or seeing images produced by AI.¹¹ Title IV, art. 52 states: "Providers shall ensure that AI systems intended to interact with natural persons are designed and developed in such a way that natural persons are informed

³ Omal Perera, *Case study: AI/AR-based Virtual Try-on for e-commerce*, Medium (Sept. 5, 2023), <https://medium.com/ascentic-technology/case-study-ai-ar-based-virtual-try-on-for-e-commerce-8d1a3d6ad6a6>.

⁴ Rebecca Cairns, *'Video games are in for quite a trip': How generative AI could radically reshape gaming*, CNN (Oct. 23, 2023, 4:40 AM), <https://www.cnn.com/world/generative-ai-video-games-spc-intl-hnk/index.html>.

⁵ Paul Sutter, *AI is already helping astronomers make incredible discoveries. Here's how*, Space (Oct. 4, 2023), <https://www.space.com/astronomy-research-ai-future#>.

⁶ Hewlett Packard, *What is AI in Finance?*, Hewlett Packard Enterprise (2023), <https://www.hpe.com/us/en/what-is/ai-in-finance.html>.

⁷ Chuck Brooks, *A Primer On Artificial Intelligence And Cybersecurity*, Forbes (Sept. 26, 2023, 8:11 PM), <https://www.forbes.com/sites/chuckbrooks/2023/09/26/a-primer-on-artificial-intelligence-and-cybersecurity/?sh=18b584f375d2>.

⁸ Rem Darbinyan, *How AI Transforms Social Media*, Forbes (Mar. 16, 2023, 8:15 AM), <https://www.forbes.com/sites/forbestechcouncil/2023/03/16/how-ai-transforms-social-media/?sh=6b680c301f30>.

⁹ Sydney Young, *The Future of Farming: Artificial Intelligence and Agriculture*, Harv. Int. Rev. (Jan. 8, 2020), <https://hir.harvard.edu/the-future-of-farming-artificial-intelligence-and-agriculture/>.

¹⁰ Orly Lobel, *The Law of AI For Good*, San Diego Legal Studies Paper 23-001 (Jan. 26, 2023) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4338862.

¹¹ EU Draft AI Act.

that they are interacting with an AI system, unless this is obvious from the circumstances and the context of use.” Quebec similarly passed a [law](#) that requires individuals to be informed when automated decision-making tools are being used.¹² Other [private](#) and public declarations about ethical AI similarly emphasize such disclosures about artificiality as a keystone of AI governance. In recent surveys, most Americans want to know when they’re interacting with an [AI](#).¹³ But why is it?

Here, in brief, I organize the reasons why policymakers and users advocate for the right to know:

1. *Dignity*

- (a) in knowing – people may hold the belief in and of itself it is essential to know who/what they interacted with or who/what made a decision pertaining to them;
- in preserving the human touch – there may be a belief that certain processes require human involvement for deontological reasons. We can call this the “AI Regulation Day in Court Lens”, even if the “court” is an administrative decisionmaker, a physician, or a customer service call center. Knowing about artificiality provides users the ability to demand reversion to human.

2. *Control* – there may be a sense that knowing the source of generated content or decisions increases control of humans, while conversely, sans disclosure, there is the risk of incremental replacement and loss of control.

3. *Trust* – knowing more about the identity of the system behind the scenes – human or AI – can give people information about whether to trust the decision/recommendation/interaction. Trust can pertain to many aspects: accuracy, safety, fairness, consistency, equity, and more. People may believe that knowing the source of a decision, recommendation, or interaction gives them better information about whether to trust – and in turn, comply with or rather override/disregard/discount – those outputs.

4. *Authenticity* – when disclosure relates to the source of AI generated photos, videos, news dissemination and reporting, knowing about the source can provide information about the veracity of the output.

¹² Samuel Adams, *Quebec’s Bill 64: The first of many privacy modernization bills in Canada?*, IAPP (Nov. 23, 2021), <https://iapp.org/news/a/quebecs-bill-64-the-first-of-many-privacy-modernization-bills-in-canada/> [<https://perma.cc/5MXY-4WCQ>].

¹³ Michelle Faverio & Alec Tyson, *What the data says about Americans’ views of artificial intelligence*, Pew Research Center (Nov. 21, 2023), <https://www.pewresearch.org/short-reads/2023/11/21/what-the-data-says-about-americans-views-of-artificial-intelligence/> [<https://perma.cc/M87F-7BLB>].

5. *Privacy*- with regards to data collection and personal information, people may hold beliefs about variance in risk of surveillance and information protection in interactions with a human versus a bot.
6. *Experiential/Artistic Integrity*- in the context of artistic works, people may feel that viewing an AI-generated image is aesthetically different than human created; that art or prose crafted by a human has greater value than when it was made by an AI.
7. *Ownership/Attribution* - the identity of the generator of a work can also give information about the ownership and intellectual property protections of the image and the legality/risk of using the image in subsequent creations.

This non-exhaustive list of concerns and goals related to the right to know is aimed to aid more deliberative discussions. What is important is to recognize that these reasons may carry different weights and have different normative implications. Indeed, disclosure may inadvertently be counterproductive from the perspective of some of the rationales. In particular, if the reason for telling users they are interacting with an AI is to increase trust and accurate evaluations of the interaction or outputs, research shows that this right to know about AI may, under certain circumstances, have inadvertent counterproductive effects.

In a [recent experiment](#) published in *Nature*, physicians received chest X-rays and diagnostic advice, some of which was inaccurate.¹⁴ While the advice was all generated by humans, some of the advice for the purpose of the experiment was labeled as generated by AI and some by human experts. In the experiment, radiologists rated the same advice as lower quality when it appeared to come from an AI system. Other studies find that when recommendations pertain to more subjective types of decisions, humans are even less likely to rely on the algorithm. This holds true even when the subjects see the algorithm outperform the human and witness the human make the same error as the [algorithm](#).¹⁵

In another [study](#) that examines the effects of replacement of a human for an automated system showed that over the course of twenty forecasting trials, people trusted automated advisors less if the automated advisor had replaced an initial human one, as opposed to having been introduced from the beginning of the

¹⁴ Susanne Gaube et al., *Do as AI Says: Susceptibility in Deployment of Clinical Decision-Aids*, 4 NPJ Digit. Med., art. 31 (2021), <https://doi.org/10.1038/s41746-021-00385-9> [<https://perma.cc/947U-M7E7>].

¹⁵ Berkeley J. Dietvorst & Soham Bharti, *People Reject Algorithms in Uncertain Decision Domains Because They Have Diminishing Sensitivity to Forecasting Error*, 31 Psych. Sci. 1302 (2020).

forecasting interactions.¹⁶ Automated advisors that replaced humans were rated as issuing lower quality advice, and human advisors that replaced automated advisors were rated as providing better quality advice. Context matters, and notably, in the financial services for example, “robot-advisors” are becoming increasingly trusted. In addition to seeing an algorithm make a mistake or issue poor advice, algorithmic aversion also increased when people had to choose between an algorithm’s forecast and their own, particularly when the people choosing had expertise in the subject they were forecasting.¹⁷

Other studies show that under certain conditions, counter-intuitively, the more an algorithm is transparent and attempts to be explainable, the more it reduces people’s ability to detect and correct model errors—perhaps because of information overload or because of a transparency trust bias—and does not appear to increase the algorithm’s acceptance.¹⁸ In one experiment, examining trust of human on moral decision-making, found that people prefer humans, who have discretion, to algorithms, which apply particular human-created fairness principles more consistently than the humans.¹⁹ Studies also show that people prefer human decision-making in inherently uncertain domains (medicine, investing).²⁰ From the perspective of giving up control, in one study, giving participants the freedom to slightly modify the algorithm made them feel more satisfied with the forecasting process, more likely to believe that the algorithm was superior, and more likely to choose to use an algorithm to make subsequent forecasts.²¹ Similarly, participants in a different study were more likely to follow algorithmic advice once their forecasts were integrated into the algorithm.²²

In a new research project on air travel with On Amir, Paul Wynns, and Alon Pereg, my collaborators and I have designed a series of experiments to examine what

¹⁶ Andrew Prael, *Algorithm Admonishment: People Distrust Automation More When Automation Replaces Humans* (working paper) (Jan. 8, 2020), <http://dx.doi.org/10.2139/ssrn.3903847> [<https://perma.cc/CGJ6-5JT3>].

¹⁷ See Logg et al. <https://dash.harvard.edu/handle/1/31677474>.

¹⁸ But see Daniel Ben David et al., *Explainable AI and Adoption of Financial Algorithmic Advisors: An Experimental Study*, AIES ’21: Proc. 2021 AAAI/ACM Conf. on AI, Ethics & Soc’y 390; J. Zerilli et al., *How transparency modulates trust in artificial intelligence*, 3 Patterns (Apr. 8, 2022), <https://doi.org/10.1016/j.patter.2022.10045> [<https://perma.cc/S7JQ-TAUV>]; Forough Poursabzi-Sangdeh et al., *Manipulating and measuring model interpretability*, Proceedings of the 2021 CHI conference on human factors in computing systems (2021).

¹⁹ Johanna Jauernig et al., *People Prefer Moral Discretion to Algorithms: Algorithm Aversion Beyond Intransparency*, 35 Phil. & Tech., art. 2 (2022), <https://doi.org/10.1007/s13347-021-00495-y> [<https://perma.cc/GKK9-UFTK>].

²⁰ Berkeley J. Dietvorst & Soham Bharti, *People Reject Algorithms in Uncertain Decision Domains Because They Have Diminishing Sensitivity to Forecasting Error*, 31 Psych. Sci. 1302 (2020).

²¹ *Id.*

²² See Kohei Kawaguchi, *When Will Workers Follow an Algorithm? A Field Experiment with a Retail Business*, 67 Mgmt. Sci. 1670 (2021).

factors contribute to trust or distrust in automation, even in situations where pilots know automation is superior. This work anticipates the next phase in aviation where completely autonomous planes will replace pilots. It also builds on initial evidence that most passengers are not entirely aware of the already robust prevalence of auto-pilot reliance in commercial aviation.

Comparing popular surveys to empirical studies reveals that public perception of the risk of unfairness of automated decision-making may simply be incorrect. For example, as I illustrate in my book *The Equality Machine*, introducing algorithmic bias in certain contexts can decrease gender, racial, and socioeconomic disparities.²³ And yet, there is a perception in public debates and people's lay assessments that a shift to automated decision-making will increase bias and inequality.²⁴ Moreover, as we saw, if a preference for human decision-makers stems from dignitarian reasons rather than outcome-driven reasons, then there inevitably may be tensions between the different rationales for distinguishing between artificial and human decision-makers. It is true that people may believe in an individual right to be in the driver's seat, literally or figuratively, or remain on the court, or in the courtroom, and that no robot should take on [tasks](#) humans have been doing for generations.²⁵ Yet, as I have argued elsewhere, policy debates need to be clear about what reasons are being invoked, what work they are doing in the analysis, and what costs we as a society are willing to pay to prioritize, for example, dignitarian harms over tangible harms, including inaccuracy, inequity, or health and safety risks.

2. Watermarking & Labeling

A distinct area in which the right to know question is becoming more salient is that of images and AI-generated artistic works. As generative AI strides deeper into creative domains, a novel ethical and legal frontier unfolds – discerning the provenance of content. AI can now conjure captivating prose, forge realistic images, and craft persuasive messages. At the same time, this has sparked a multitude of concerns, from copyright infringement to manipulated news narratives and blurring truth into fiction. Enter the burgeoning realm of AI content identification, where regulatory and private solutions grapple with methods of unveiling the algorithmic hand behind the creative brushstroke. Regulatory regimes around the

²³ Orly Lobel, *The Equality Machine* (2022).

²⁴ Nicholas Scurich & Daniel A. Krauss, *Public's Views of Risk Assessment Algorithms and Pretrial Decision Making*, 26 Psych., Pub. Pol'y and L. 1 (2020); see also Theo Araujo et al., *In AI We Trust? Perceptions About Automated Decision-Making by Artificial Intelligence*, 35 AI & Soc'y 611 (2020) (describing perceptions and realities about the risks, trustworthiness, and fairness of AI reveal gaps).

²⁵ Relatedly, human decision-makers, such as referees in athletic matches or judges in the courtroom, serve not only in handing down tangible decisions but also play a performative role. See Michael J. Madison, *Fair Play: Notes on the Algorithmic Soccer Referee*, 23 Vand. J. Ent. & Tech. L. 341 (2021) <https://scholarship.law.vanderbilt.edu/jetlaw/vol23/iss2/4> [<https://perma.cc/SX3J-X3TQ>] (relating to the algorithmic trust question discussed *infra* section 3C).

world are contemplating mandatory disclosures or marking the source of image generation. Here too, the reasons for source identification are varied. Visibly labeling an AI-generated image as such can safeguard consumers from deepfakes and deceptive marketing. A separate concern is that of ownership and intellectual property rights.

President Biden’s October 2023 [Executive Order](#) calls upon federal agencies to create recommendations to agencies for “reasonable steps to watermark or otherwise label output from generative AI.”²⁶ China has banned AI-generated images without watermarks. Watermarking techniques, akin to invisible digital signatures, are being developed to embed imperceptible markers within AI-generated content. These markers, later unearthed by dedicated software, serve as a digital fingerprint, revealing the algorithmic origin of the work. The [purpose](#) of such invisible watermarking is that of IP rights protection: to exclude copyrighted materials from training data of generative AI, thus avoiding the risk of AI memorizing and copying images, thereby infringing intellectual property.²⁷ Companies are increasingly introducing innovative methods like watermarking and digital signatures to identify AI-generated content. For instance, some AI content creation platforms embed subtle, unique markers or metadata in the content they generate. Fingerprinting technology, on the other hand, refers to software programs that analyze stylistic quirks and patterns within content, statistically determining the likelihood of AI authorship.

One potential solution lies in implementing targeted regulations focused on specific high-risk domains, such as political advertising or news media. In the context of deepfakes—especially grossly misleading images depicting actual people—the goal of labeling is to prevent [misinformation](#).²⁸ In some contexts, like election campaigns, the need for such tracking the source of content may be especially important due to the potential for AI-generated content to spread misinformation and influence public opinion. The importance of mitigating misinformation is paramount in considering the development of watermarking and disclosure tools. This targeted approach could balance and minimize the potential stifling of innovation in broader creative fields while still safeguarding against the dissemination of misleading or harmful content. From a governance perspective, promoting education, voluntary

²⁶ The White House, *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, (Jan. 15. 12:00 AM), <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>.

²⁷ Sag, Matthew. “Copyright Safety for Generative AI.” *Forthcoming in the Houston Law Review* (2023): pp. 343.

²⁸ Jack Langa, Deepfakes, Real Consequences: Crafting Legislation to Combat Threats Posed by Deepfakes, 101 B.U. L. REV. 761 (2021): pp. 789.

disclosure, and fostering a culture of transparency where artists openly acknowledge the use of AI tools can be an alternative to harder forms of command-and-control regulation and empower creators and audiences alike.

In addition to watermarking tools, there are other alternative tools to distinguish AI-generated content from content created by humans. “Data poisoning” tools insert invisible changes to the pixels of art to disrupt the training of AI models if the work is scraped into machine learning systems. These data poisoning tools are different from watermarking tools in that the pixels are invisible to the human eye. Here too the technology becomes something of a cat and mouse race: new techniques are developed to bypass poisoning, leading to a push to develop other methods of poisoning.

3. Conclusion

Is there indeed a broad principled need to distinguish between human-generated and AI-generated content? The answer, as this essay has sought to show, is no. However, the answer should depend on the goals. Ultimately, bridging the gap between human-machine interactions and trust requires understanding when and where disclosure about artificiality paves the way for achieving our individual and social goals. In this short essay, I aimed to unpack some of the rich context surrounding the questions about knowing that content, decisions, or interactions were generated by AI. This varied and complex set of considerations and implications point to the need for more deliberative and nuanced analyses of the right to know about the involvement of AI. It points to a further need for more robust behavioral research on how to establish not only trustworthy AI, but also human rationality and trust *in* AI. Such research and empirical findings should, in turn, cultivate more modernized and holistic public policies about AI deployment.

AI-generated content is a multifaceted challenge that requires a collaborative approach between governments and private entities. Private solutions, such as watermarking and AI detection tools, provide practical ways to distinguish between human and AI-generated content for specific purposes.

Orly Lobel

Warren Distinguished Professor of Law, and the founding director of the Center for Employment and Labor Policy (CELP) at University of San Diego.

Citation: Orly Lobel, *Do We Need to Know What Is Artificial? Unpacking Disclosure & Generating Trust in an Era of Algorithmic Action*, Dynamics of Generative AI (ed. Thibault Schrepel & Volker Stocker), Network Law Review, Winter 2024..