Harvard Journal of Law & Technology
Volume 31, Number 1 Fall 2017

ARTIFICIAL INTELLIGENCE AS A DIGITAL PRIVACY PROTECTOR

*Andrea Scripa**

TABLE OF CONTENTS

I. INTRODUCTION

Artificial Intelligence ("AI") and its potential to effect dramatic social and scientific change has long been a subject of our collective fascination. Popular culture has simultaneously explored the potential for AI to resolve some of our greatest scientific challenges, and to lead to an uprising of self-aware robot assassins.[1] Contemporary challenges in AI are less dramatic, but equally difficult to navigate. Machine learning has supplemented quintessential artificial intelligence activities such as computer vision, natural language processing, and navigation. Rather

[*] Harvard Law School, J.D. 2017; Duke University, B.S. in Computer Science & B.A. in Linguistics, 2012. I would like to thank Professor Urs Gasser, who advised the paper that led to this Note, Tina Chao for her helpful comments as Article Editor, and the wonderful editors at the Harvard Journal of Law and Technology for their diligence and support.

[1] *Compare* TRANSCENDENCE (Alcon Entertainment 2014) *with* THE TERMINATOR (Hemdale 1984). *See also* Joshua Ostroff, *NASA Scientist: Artificial Intelligence 'Could Solve All The World's Problems' (If It Doesn't Terminate Us)*, HUFFINGTON POST (June 26, 2015, 12:59 PM), http://www.huffingtonpost.ca/2015/06/26/artificial-intelligence-ai-richard-terrile-nasa_n_7654630.html [https://perma.cc/2DTU-K6UU]; Lucy Draper, *Could Artificial Intelligence Kill Us Off?*, NEWSWEEK (June 24, 2015, 7:38 AM), http://europe.newsweek.com/could-artificial-intelligence-kill-us-off-329208 [https://perma.cc/Z398-MWMP] (including quotes from Elon Musk and Stephen Hawking that AI could well become an existential threat).

than emulating human abilities, machine learning endeavors to accomplish tasks that would otherwise be impracticable for humans to perform alone, primarily by processing giant swaths of data and extracting hidden insights. Machine learning has proven useful not only in improving performance in traditional AI applications, but also in countless new applications. These include expanding the coverage insurance companies can provide, personalizing advertising and recommendation engines, and outperforming doctors in specialized medical diagnoses.[2]

Such activities have led to the realization that many knowledge gaps in computerized systems can be filled more easily by providing the system with large amounts of data than by expending immense resources trying to perfect the underlying algorithm.[3] As a result, machine learning derives much of its power from the "Big Data Revolution": our rapidly increasing ability to collect, store, and process large quantities of detailed information.[4]

The widespread use of AI and machine learning (treated collectively as AI in this Note) has obvious implications for data privacy, as well as for privacy-related values such as autonomy, due process, and equality.[5] For example, AI has the power to reveal information that would not be obvious to a human evaluating a dataset unassisted, including sensitive information that was never disclosed by the subject. Target's now-famous "pregnancy prediction score" was able to correctly guess that a teenage girl was pregnant based on her purchases of a handful of common products.[6] Her father became enraged that Target sent the teen baby-related advertisements, only to find out later that she was, in fact, pregnant.[7] Other predictive scores have the potential to be not only creepy, but also discriminatory and unjust. Algorithms that determine

[2] *See* Eric Brat et al., *Bringing Big Data to Life: Four Opportunities for Insurers*, BCG PERSPECTIVES (July 17, 2014), https://www.bcgperspectives.com/content/articles/insurance_digital_economy_Bringing_big_data_life/ [https://perma.cc/A28F-AFGB]; Matthew Hutson, *Self-Taught Artificial Intelligence Beats Doctors at Predicting Heart Attacks*, SCIENCE (Apr. 14, 2017, 3:30 PM), http://www.sciencemag.org/news/2017/04/self-taught-artificial-intelligence-beats-doctors-predicting-heart-attacks [https://perma.cc/3YH9-UYYW].

[3] *See* STUART RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 27–28 (3d ed. 2010).

[4] *See The Big Data Revolution*, IEDP (Feb. 27, 2012), http://www.iedp.com/articles/the-big-data-revolution/ [https://perma.cc/FA4E-6Z3M]. In addition to recent increases in memory and storage capacity, many Big Data applications have been made possible by cloud computing and distributed analytics technologies. *See* Sheri Pan, Note, *Get To Know Me*, 30 HARV. J.L. & TECH. 239, 246 (2016) (referencing database sharding, NoSQL, MapReduce, Yarn, and Hadoop as examples).

[5] *See* URS GASSER, THE FUTURE OF DIGITAL PRIVACY: A NAVIGATION AID (forthcoming 2018).

[6] *See* Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. TIMES (Feb. 16, 2012), http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html (last visited Dec. 20, 2017).

[7] *Id.*

who should receive a mortgage or what an individual's propensity for future criminal activity might be often use data that reflects systemic disparities in society, and thus might perpetuate or worsen unequal treatment.[8] Furthermore, no one has yet determined how to adequately audit machine learning algorithms for bias, as they rely on extremely large datasets,[9] and are both nondeterministic and mathematically complex.[10] Not knowing how to evaluate the reliability and fairness of an algorithm makes it difficult to hold its proponents accountable for the outcomes it produces and, if necessary, to have its use discontinued. This prospect is especially disconcerting given that algorithms on everyday platforms like Facebook, Netflix, and Amazon can influence our personalities and preferences, and sometimes make decisions for us without any human input at all.[11]

Much remains to be done to ensure that existing uses of AI are subject to appropriate scrutiny and regulation so as not to undermine privacy and other important values. This Note, however, explores the ways in which AI and related computational methods might be used to *enhance* protections for personal privacy, either by mitigating AI's own negative effects or by addressing privacy concerns caused by other trends. Part II outlines four significant ways in which AI might be used to bolster privacy protections, while Part III proposes strategies for reinforcing these technological solutions with legal measures or private conduct. Part IV concludes.

II. Privacy-Enhancing AI Technologies

This Part will discuss several ways in which AI can be used to enhance personal privacy and will analyze the advantages and limitations of each method. It will begin with two relatively established AI-based privacy enhancing technologies ("PETs") – differential privacy and federated learning – and then will proceed to evaluate more speculative uses of AI as auditors and guardians. The final Section proposes that AI might even be used to define privacy itself.

[8] *See generally* Cathy O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (2016).

[9] Jonathan Cohn, *The Robot Will See You Now*, The Atlantic (March 2013), https://www.theatlantic.com/magazine/archive/2013/03/the-robot-will-see-you-now/309216 [https://perma.cc/4S95-MRNU] (announcing that IBM's Watson can process up to sixty million pages of text per second).

[10] *See generally* Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information (2015).

[11] *See* Gasser, *supra* note 5.

A. Differential Privacy

Differential privacy is a field pioneered by researchers at Microsoft and Apple, alongside a handful of academics. The animating principle behind differential privacy, as articulated by its original proponent Cynthia Dwork, is that responses to dataset queries should not provide enough information to identify any individual included in the dataset.[12] Differential privacy is ultimately a mathematical definition of privacy that considers whether a particular person's data has a significant impact on the answer to a dataset query; if it does not, then the data will not identify the person it describes.[13] The identifiability of information is (as we have undoubtedly discovered)[14] not a binary question, but a probabilistic one. How much of an impact the data must have on the query to be excluded — and by extension how likely it is that a query would lead to personal identification — depends on a "privacy budget" set by the holder of the data, which defines how much information leakage is considered acceptable.[15]

Setting an appropriate privacy budget is therefore crucial to the proper use of differential privacy techniques. And because of the way that differential privacy works, there is an inherent tradeoff between the level of privacy afforded to data subjects and the accuracy of the query results. This is because differential privacy is performed primarily by injecting noise (randomness) into a dataset in such a way that the outputs or conclusions generated by the data are minimally impacted while privacy protection is enhanced.[16] The amount of noise introduced will depend on the specified amount of acceptable data leakage and the way the data will be used. Just as data leakage will never reach zero, neither will the amount of error introduced by the noise.

Apple has developed more sophisticated differential privacy techniques that incorporate hashing and subsampling into its methodology as

[12] *See* Cynthia Dwork*, Differential Privacy,* 33 INT'L COLLOQUIUM ON AUTOMATA, LANGUAGES AND PROGRAMMING 1 (2006).

[13] Matthew Green, *What Is Differential Privacy?*, A FEW THOUGHTS ON CRYPTOGRAPHIC ENGINEERING (June 15, 2016), https://blog.cryptographyengineering.com/2016/06/15/what-is-differential-privacy/ [https://perma.cc/73YU-RKJZ].

[14] For example, Netflix's publicly released viewing dataset for an algorithmic design contest turned out to be insufficiently anonymized because researchers discovered that the dataset could be used to re-identify certain viewers when combined with publicly-available data. This led to inquiries by the FTC and a California class-action lawsuit against Netflix. *See* Andrew Chin & Anne Klinefelter, *Differential Privacy as a Response to the Reidentification Threat: The Facebook Advertiser Case Study*, 90 N.C. L. REV. 1417, 1424 (2012). In another case, Latanya Sweeney published a study in which she merged supposedly anonymized Massachusetts worker hospital records with easily acquired voter registration records, and found she was able to identify the health records of then-Governor William Weld; she later published "a broader study finding that 87% of the 1990 U.S. Census population could be identified using only gender, zip code, and full date of birth." *Id.* at 1425.

[15] Green, *supra* note 13.

[16] *Id.*

well.[17] Subsampling, for example, is the straightforward practice of processing a small selection of data rather than the entire dataset. Hashing involves transforming data into a seemingly random set of values in a mathematically deterministic but difficult-to-reverse way.[18] Because the same input always yields the same (hopefully unique) output, storing hashed data allows data collectors to track similarities in returned values without having to store the original information. Hashing is often used by itself to anonymize sensitive data;[19] however, it leaves significant privacy issues unresolved. For example, the input data could still be derived by a determined adversary through inefficient methods like guess-and-check, or could be uncovered by examining related data that was not hashed. The use of hashing in the context of differential privacy may help Apple construct privacy filters that both hash and randomize responses before they are sent back to the company.[20] This prevents Apple from ever having to store and secure the raw underlying data, and provides additional guarantees against re-identification that hashing alone does not offer.

Apple has already implemented differential privacy in iOS 10, which will randomize device data before sending it back to Apple and will limit the amount of data that can be collected from any one user.[21] Apple has indicated that the differentially private data it collects will be used to improve its keyboard, rank deep-linked search results in Spotlight searches, and recommend actions based on information entered into the Notes app.[22] Other likely applications include determining the popularity of products, emojis, or news topics, and facilitating troubleshooting for common iOS bugs.[23]

Other promising applications of differential privacy involve collecting raw user data, and then fabricating entirely new datasets with the same mathematical properties as the original, such that none of the original data which corresponds to a real person is saved.[24] This method

[17] Apple, *Apple— WWDC 2016 Keynote*, YOUTUBE (June 14, 2016), https://www.youtube.com/watch?v=n5jXg_NNiCA (last visited Dec. 20, 2017) (showing at 102:08–102:22 Apple SVP Craig Federighi describe Apple's differential privacy practices as "us[ing] hashing, subsampling, and noise injection to enable . . . crowdsourced learning while keeping the information of each individual user completely private").

[18] *Id.*

[19] *See* Kate Conger & Natasha Lomas, *What Apple's Differential Privacy Means for Your Data and the Future of Machine Learning*, TECHCRUNCH (June 14, 2016), https://techcrunch.com/2016/06/14/differential-privacy/ [https://perma.cc/L8JK-EK6S].

[20] *See* Green, *supra* note 13.

[21] *See* Conger & Lomas, *supra* note 19.

[22] *See id.*

[23] Gennie Gebhart et al., *Facial Recognition, Differential Privacy, and Trade-Offs in Apple's Latest OS Releases*, ELEC. FRONTIER FOUND. (Sept. 27, 2016), https://www.eff.org/deeplinks/2016/09/facial-recognition-differential-privacy-and-trade-offs-apples-latest-os-releases [https://perma.cc/P6YC-6EWA].

[24] *See* Ira S. Rubenstein & Woodrow Hartzog, *Anonymization and Risk*, 91 WASH. L. REV. 703, 718–19 (2016).

could prove promising for data collectors who share datasets with third parties, each of whom investigates different questions with different levels of privacy risk. It is also helpful when the raw data is particularly sensitive, as is the case with health data.

Despite providing many advantages over previous privacy-enhancing techniques, differential privacy has its share of limitations. Chiefly, differential privacy focuses on preventing the re-identification of data subjects from information revealed by the dataset, but it cannot prevent information from being revealed due to a requester's prior knowledge.[25] Furthermore, the use of differential privacy typically requires that a number of conditions be met: namely, that the total amount of data in the dataset be relatively large, that the use of the data be able to tolerate some distortion, that the lower and upper bounds of numerical answers be known, and that the outliers in a dataset not be particularly important.[26] Finally, as indicated previously, the more detailed the information one intends to request, the less reliable the results will need to be in order to limit privacy leakage, and once the database has leaked as much information as the calculations determine is safe, "you can't keep going."[27] Differential privacy is nonetheless a substantial improvement over previous measures designed to prevent re-identification when it is available as an option.

B. Federated Learning

Federated learning, a process recently developed by Google,[28] allows a centralized machine learning model to receive feedback from users without storing their individual data in the cloud. The individual's device still collects data to improve the model, but instead of sending up the raw data, it determines the changes that should be made to the model

[25] For example, a dedicated sleuth who has collected data about a person from multiple sources may combine it to uncover a full picture that would have been obfuscated if requested from a single, differentially private dataset. *See* Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 U. COLO. L. REV. 1117, 1137 (2013) ("[I]t is always theoretically possible that any information revealed by a data set is the missing link that the adversary needs to breach someone's privacy.").

[26] Chin & Klinefelter, *supra* note 14, at 1417, 1448–52.

[27] Green, *supra* note 13. Data leakage can also be accelerated if the same question is answered multiple times, with different degrees of noise injected into the answer each time. *Id.*; *see also* Wu, *supra* note 25, at 1140 ("[G]etting answers to too many questions about arbitrary sets of individuals in a sensitive data set allows an adversary to reconstruct virtually the entire data set, even if the answers he or she gets are quite noisy.").

[28] *See* Brendan McMahan & Daniel Ramage, *Federated Learning: Collaborative Machine Learning Without Centralized Training Data*, GOOGLE RES. BLOG (Apr. 6, 2017), https://research.googleblog.com/2017/04/federated-learning-collaborative.html [https://perma.cc/4W58-23NY]; *see also* James Vincent, *Google Is Testing a New Way of Training Its AI Algorithms Directly on Your Phone*, THE VERGE (Apr. 10, 2017), http://www.theverge.com/2017/4/10/15241492/google-ai-user-data-federated-learning [https://perma.cc/RHX4-HG9N].

locally and then sends a "small focused update" to the cloud, where the update is averaged with other updates to improve the model.[29] Neither individual data nor individual updates need ever be stored in the cloud — only the averaged 'meta-update' is retained.[30] Federated learning therefore allows Google to develop more privacy-protective machine learning models that rely on individualized data, despite the fact that the data the model consumes is unevenly distributed across millions of devices, which are available for training only intermittently.[31] Google overcame the latency problem associated with receiving updates directly from customers' Androids by crafting updates of particularly high computational quality.[32] That way, fewer of them would be needed to optimize the model.

Federated learning marks a significant breakthrough in many fields, including artificial intelligence, compression, and cryptography. It differs from differential privacy in that it does not require the same privacy/data quality tradeoff, but this comes at the cost of more challenging data collection logistics. It also prevents Google from putting the data it collects to multiple uses, as the updates that are generated are designed to be plugged directly into a specific machine learning algorithm.

Google has already started using federated learning to improve the word recommendations made by the Android keyboard, Gboard, and envisions additional applications in language modeling and photo rankings.[33] That said, federated learning is not a silver bullet. It cannot be used with algorithms that require labeling raw data, and of course will not be necessary when the training data is already stored in the cloud, as is true for the Gmail data used in spam filtering.[34] Google's researchers are now looking to extend federated learning to be compatible with additional machine learning algorithms, such as neural networks, that are state-of-the-art in important application areas.[35] Eventually, federated learning might be applied to any problem that relies on data collected by devices, which holds great promise for securing personal data collected by the Internet of Things.

[29] McMahan & Ramage, *supra* note 28.

[30] *Id*. Even then, Google has imposed a limitation called the Secure Aggregation protocol, which "uses cryptographic techniques so a coordinating server can only decrypt the average update if 100s or 1000s of users have participated"; this prevents inspection of a meta-update that contains data from too few users. *Id*.

[31] *Id.*

[32] *See id.*

[33] *See id.*

[34] *Id.*

[35] Jakub Konečný et al., *Federated Optimization: Distributed Machine Learning for On-Device Intelligence*, ARXIV 26 (Oct. 11, 2016), https://arxiv.org/pdf/1610.02527.pdf [https://perma.cc/YHM3-PA5H].

C. AI Auditors and Personal Guardians

The idea of AI-based agents is an old one, and many companies have adopted automated customer service agents to provide phone and chat support to consumers. But the idea of using these agents to protect privacy is relatively new. Last year, MIT researchers developed a platform called AI2, which is capable of identifying suspicious activity from approximately 85% of cyberattacks by combing through log data and identifying suspicious activity.[36] Importantly, AI2 is able to process incoming data in real time and can generate and refine new models in response to feedback in as little as a few hours.[37] This technology has great potential to improve the security of consumers' data because it can flag only the most useful information for security experts to review, ensuring quicker responses to data breaches. The importance of developing intelligent alert systems for networks and databases will only increase with the continued growth of Big Data and the Internet of Things.

The concept of an AI auditor could be extended to other privacy-related areas as well, such as to prevent re-identification or to identify algorithmic outcomes that are unfair and discriminatory. In addition to implementing differential privacy measures on a dataset, for example, an AI auditor could monitor the use of the differentially private dataset to ensure that it is used only for its intended purpose, and to prevent further use of the data once the privacy budget has been used up. Essentially, the proposal is to develop AI intended to guard other AI in its management of personal data. Technology experts have foreseen this general application of AI, in fact, and have suggested that it may become crucial to AI's continued widespread use, as humans cannot adequately monitor highly complex and constantly changing models on their own.[38]

Beyond serving as a mere auditor, AI could be assigned to the role of guardian, with some experts requesting "algorithmic angels" to represent our interests as we navigate a world full of automated systems.[39] An AI guardian would not only monitor for common problems in AI systems, but also advocate for its master's interests. AI guardians could act as countermeasures to over-personalization and insidious algorithmic manipulation, collaborating with other technologies to better accommo-

[36] Adam Conner-Simons, *System Predicts 85 Percent of Cyber-Attacks Using Input from Human Experts*, CSAIL (Apr. 18, 2016), http://www.csail.mit.edu/System_predicts_85_percent_of_cyber_attacks_using_input_from_human_experts [https://perma.cc/LMA3-SX7B].

[37] *See id.*

[38] *See* Amitai Etzioni & Oren Etzioni, *Keeping AI Legal*, 19 VAND. J. ENT. & TECH. L. 133, 137–41 (2016).

[39] Jarno M. Koponen, *We Need Algorithmic Angels*, TECHCRUNCH (Apr. 18, 2015), https://techcrunch.com/2015/04/18/we-need-algorithmic-angels/ [https://perma.cc/36JM-6ZEZ].

date a user's preferences and ensure that the user's decisional autonomy is not being subtly undermined.[40]

Companies like Uber have already started to algorithmically exploit cognitive biases in their workers to get them to serve the company's ends at minimal cost. Much like Netflix's continuous streaming of shows, which has been found to encourage binge-watching, Uber's app keeps drivers constantly busy and automatically loads a driver's next ride as the current ride concludes.[41] This raises the mental decision cost to the driver to stop driving, which steers workers to continue accepting rides for longer than they otherwise would.[42] Uber also uses drivers' tendency to set earnings goals to keep them driving longer, informing them of arbitrary sums that they've almost earned.[43] This practice of displaying goals perceived to be just beyond a subject's grasp is part of what makes gambling so addictive. Electronic slot machines, for example, entice players to continue gaming by providing periodic positive reinforcement, such as small wins and near-misses.[44] Although these near-misses are no closer to a win than any other losing result, gamblers keep playing due to the irrational feeling that they *almost* won.[45] AI guardians could counteract these effects with carefully timed locks or reminders that encourage better decision-making. With the right algorithmic pushback, the same technology that allows a company to manipulate its workforce could instead be used to create more stability and convenience for workers.

Other sources have suggested using AI guardians to negotiate data sharing[46] or privacy policy terms.[47] A project called Customer Com-

[40] *See id.*

[41] *See* Noam Scheiber, *How Uber Uses Psychological Tricks to Push Its Drivers' Buttons*, N.Y. TIMES (Apr. 2, 2017), https://www.nytimes.com/interactive/2017/04/02/technology/uber-drivers-psychological-tricks.html (last visited Dec. 20, 2017). *See generally* Ryan Calo & Alex Rosenblat, *The Taking Economy: Uber, Information, and Power*, 117 COLUM. L. REV., (forthcoming 2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2929643 (last visited Dec. 20, 2017).

[42] *See id.*

[43] Scheiber, *supra* note 41. A similar experiment performed by Lyft found that drivers could be even more easily incentivized to drive at certain times by manipulating their loss aversion: showing them how much they stood to lose by not changing their schedules, as opposed to how much they would gain by doing so. *Id.*

[44] *See The Almost-Winning Addiction*, ECONOMIST (May 6, 2010) http://www.economist.com/node/16056339 [https://perma.cc/6ZJ2-UR5X].

[45] *See id.*

[46] *See* IEEE GLOBAL INITIATIVE FOR ETHICAL CONSIDERATIONS IN ARTIFICIAL INTELLIGENCE AND AUTONOMOUS SYSTEMS, ETHICALLY ALIGNED DESIGN 67 (2016), https://standards.ieee.org/develop/indconn/ec/ead_personal_data.pdf [https://perma.cc/8VQV-22Z3] ("The guardian could serve as an educator and negotiator on behalf of its user by suggesting how requested data could be combined with other data that has already been provided . . . [T]he guardian could negotiate conditions for sharing data and could include payment to the user as a term . . . .").

mons is developing an automated platform that enables web users to communicate the minimum terms they would be willing to agree to in order to use a website.[48] This tool would standardize certain privacy benchmarks that users could collectivize around, permitting users to demand certain standards as a group that companies otherwise would not need to offer. In addition to the collectivization feature, Customer Commons empowers users by automating a contracting process that is currently too varied and time-consuming for them to monitor without assistance.[49] It converts a one-sided negotiation to a more two-sided one.

Lastly, AI guardians could protect privacy in the consumption of products in addition to services. Aggregating consumers into product-buying groups helps consumers obscure their personal preferences and counteracts highly detailed algorithmic price discrimination schemes.[50] Currently, online retailers track or purchase information about consumers' demographic background and online behavior to estimate the highest price they could be expected to pay for a product.[51] Just knowing an individual's zip code, paired with the current weather conditions of that area, could indicate an increased proclivity to buy jackets and antidepressants if it is raining, or convertibles and antihistamines if it is sunny. Knowing an individual's recent searches and interests adds even greater predictive power. Such detailed modeling allows for more precise price discrimination than ever before.[52] Algorithmic buying groups are one possible solution to the subtle price manipulation, carefully placed online advertisements, and optimally timed incentives that have slowly become more and more pervasive. AI can reallocate some of consumers' recently-lost economic surplus back to them, ensuring that society strikes a healthy balance between the interests of companies and con-

[47] *Cf. Customer Commons and User Submitted Terms,* CUSTOMER COMMONS (Oct. 27, 2014), http://customercommons.org/2014/10/27/customer-commons-and-user-submitted-terms/ [https://perma.cc/PC3P-BKCX].

[48] *Id.*

[49] This application is analogous to the emerging use of bots in renegotiating rates, resolving billing discrepancies, and terminating unwanted subscriptions with telecom providers: an industry with notoriously — and perhaps intentionally — bad customer service. Last year, a startup called Trim unveiled an AI extension for Google Chrome that would renegotiate users' cable bills with Comcast via online chat. *See* Megan Farokhmanesh, *Stop Arguing with Comcast and Let This Bot Negotiate for You*, THE VERGE (Nov. 17, 2016, 10:16 AM), https://www.theverge.com/2016/11/17/13656264/comcast-bill-negotiator-bot-argue-money-customer-service [https://perma.cc/VYM4-EE8W]. These bots serve as repeat players in negotiations against telecom providers, gathering data about consumers' typical rates and negotiation strategies, and minimizing the tax on consumers' time, energy, and patience.

[50] *See* Michal S. Gal & Niva Elkin-Koren, *Algorithmic Consumers*, 30 HARV. J.L. & TECH. 309, 331–34 (2017).

[51] *See* Maurice Stucke, *Virtual Competition: The Promise and Perils of the Algorithm-Driven Economy* (Mar. 28, 2017), https://cyber.harvard.edu/events/luncheons/2017/03/Stucke [https://perma.cc/KWN5-R3ZC].

[52] *Id.*

sumers, and protecting personal autonomy in economic decision-making.

D. Using AI to Define Privacy

Finally, AI may help protect privacy by simply helping us define what privacy is. Scholars have struggled to articulate a unifying theory of privacy, which has alternatingly been defined as control over information,[53] as our accessibility to others,[54] and as contextual integrity.[55] The theory of privacy as contextual integrity appears to be the most sophisticated theory: it acknowledges that privacy itself is a social construct that is influenced by other social norms. People have different expectations of privacy in different spheres of their lives because each of these spheres is governed by a different set of rules. Privacy as contextual integrity is not a fully defined theory but — as its originator Helen Nissenbaum emphasizes — a "benchmark."[56] Contextual integrity requires that we look to our collective social norms to define what privacy means, and it is for this reason that the theory would readily lend itself to refinement through AI.

By examining data on what is and is not considered a privacy violation in various circumstances, AI methods (namely, classification algorithms) could help uncover insights about the most salient contextual features and informational norms that govern privacy determinations. A well-trained model could one day even outperform individual people in predicting what our collective wisdom would consider private.[57] This information could then be used to inform privacy-related policy decisions — an area that frequently remains paralyzed due to a lack of consensus on what privacy entails.

[53] *See* ALAN F. WESTIN, PRIVACY AND FREEDOM 7 (1967) ("Privacy is the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others.").

[54] *See* Ruth Gavison, *Privacy and the Limits of Law*, 89 YALE L.J. 421, 423 (1980) ("Our interest in privacy . . . is related to our concern over our accessibility to others: the extent to which we are known to others, the extent to which others have physical access to us, and the extent to which we are the subject of others' attention.").

[55] *See* Helen Nissenbaum, *Privacy as Contextual Integrity*, 79 WASH. L. REV. 119, 138 (2004) ("[T]he benchmark of privacy is contextual integrity; that in any given situation, a complaint that privacy has been violated is sound in the event that one or the other types of informational norms has been transgressed.").

[56] *Id.*

[57] Of course, AI models must be trained on data, which raises the question of where this data would come from. There are several possibilities. It could be collected through experiments or surveys using methods like the "Moral Machine" questionnaire at the MIT Media Lab, which queries visitors' perspectives on the values that should be programmed into autonomous vehicles. *See* MORAL MACHINE, MASS. INST. OF TECH., http://moralmachine.mit.edu/ [https://perma.cc/C5DT-VNJD]. Alternatively, the data could be scraped from the results of legal cases, or repurposed from past studies in the social sciences on how people conceptualize and respond to privacy issues.

Similarly, AI would be useful not just for developing a more robust theory of privacy, but also for determining when a violation of one's privacy has caused harm. U.S. courts have struggled to determine what constitutes a cognizable privacy harm. Not all violations of privacy lead one to suffer a privacy-related harm, and not all privacy-related harms have a legal remedy. Cases considering the requirements for Article III standing readily illustrate this principle. In *Clapper v. Amnesty International USA*,[58] the Supreme Court held that Article III's "injury in fact" requirement does not encompass internal or subjective privacy-related harms.[59] Three years later, the Court concluded in *Spokeo v. Robins*[60] that a statutory violation likewise would not provide Article III standing absent some injury that is "concrete and particularized."[61] The Court did not spell out exactly what the concreteness requirement entails, though it conceded that harms need not immediately translate into an injury if there is a significant risk of a real harm occurring later.[62] Large-scale breaches of Equifax and national health insurer CareFirst during the writing of this Note have further intensified the debate around standing, and the Court has been asked to clarify the standard it articulated in *Spokeo*.[63]

All of this makes paramount the question of how much risk of harm is required for a breach of privacy when an observable harm has not yet taken place. AI techniques could provide courts with a more concrete idea of how likely an individual is to suffer a harm that has not yet occurred, in the event that her information was subject to a security vulnerability or unlawfully disclosed to a third party. If provided information about the event and about similar events that have occurred in the past, an algorithm could estimate the probability of an individual's information being used against her, as well as the likely magnitude of any financial losses. This practice would allow courts to engage in more evidence-based decision-making around uncertain privacy harms, and has

[58] 568 U.S. 398 (2013).

[59] *Id.* at 416 ("[R]espondents cannot manufacture standing merely by inflicting harm on themselves based on their fears of hypothetical future harm that is not certainly impending.").

[60] 136 S. Ct. 1540 (2016).

[61] *Id.* at 1548.

[62] *See id.* at 1549–50.

[63] *See* Amul Kalia & Cindy Cohn, *Will the Equifax Data Breach Finally Spur the Courts (and Lawmakers) to Recognize Data Harms?*, ELECTRONIC FRONTIER FOUND. (Sept. 26, 2017), https://www.eff.org/deeplinks/2017/09/will-equifax-data-breach-finally-spur-courts-and-lawmakers-recognize-data-harms [https://perma.cc/7MXW-ENWH]; Amy Aixi Zhang, *Attias v. CareFirst: CareFirst Petitions for Cert to Decide Standard of Harm in Data Breach Cases*, JOLT DIGEST (Nov. 13, 2017), http://jolt.law.harvard.edu/digest/attias-v-carefirst-carefirst-petitions-for-cert-to-decide-standard-of-harm-in-data-breach-cases [https://perma.cc/TAV3-GWFA]. Spokeo itself recently filed a second writ of certiorari, arguing that the standing bar the Court set in 2016 has led to widespread confusion and inconsistent lower court decisions. *See* Allison Grande, *Spokeo Wants Justices to Revisit Last Year's Standing Ruling*, LAW360 (Dec. 13, 2017, 10:50 PM), https://www.law360.com/cybersecurity-privacy/articles/994507/spokeo-wants-justices-to-revisit-last-year-s-standing-ruling (last visited Dec. 20, 2017).

precedent in bail and sentencing decisions, which also often rely on algorithmic calculations.[64] AI seems particularly well-suited to the context of estimating privacy harms because doing so requires reasoning on abstract, uncertain, and seemingly rare events, a task that humans perform poorly.[65]

Using AI to help us define both a theory of privacy and privacy-related harms is not without its risks, however. The most obvious risks, articulated above, are that algorithms can be provided with biased data, that they are difficult to monitor and understand, or that they may give results that are hard to interpret. This class of common problems must be addressed more comprehensively before AI permeates more areas of everyday life. But beyond these issues of algorithmic accountability is a more philosophical problem: what will happen to our social norms and our ability to make moral decisions if we allow AI to make these decisions for us?

At some point, using AI to define privacy could turn into an activity that is prescriptive rather than merely descriptive. People may become accustomed to allowing AI to resolve privacy challenges and cease to question its conclusions, thereby ossifying existing privacy conventions and making compliance with the algorithm a norm unto itself. Scholars have argued, for instance, that the use of digital rights management ("DRM") technologies in digital media has led to a breakdown in intellectual property ("IP") law.[66] As consumers have grown more accepting of DRM, companies have become emboldened to undermine the economic and political bargains that lie at the root of IP, including the fair use doctrine.[67] It is important that AI remain just one tool we use to help us understand ourselves. Incorporating AI into policymaking should not absolve us of responsibility to engage in our own reasoning about important normative questions.

III. REINFORCING THE USE OF AI PETS

The four PETs outlined above are promising, but must be reinforced with other private and legal measures to ensure they are actually used, and used appropriately. In particular, technological approaches are often

[64] *See* Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/PYL5-KKWF].

[65] *See generally* NASSIM NICHOLAS TALEB, THE BLACK SWAN: THE IMPACT OF THE HIGHLY IMPROBABLE (2010).

[66] *See generally* Christopher May, *Digital Rights Management and the Breakdown of Social Norms*, 8 FIRST MONDAY (2003).

[67] *Id*. *But see* Tom W. Bell, *Fair Use Vs. Fared Use: The Impact of Automated Rights Management on Copyright's Fair Use Doctrine*, 76 N.C. L. REV. 557, 578–600, 619 (1998) (arguing that the rise of DRM may create a more, not less, equitable environment for digital media creators and consumers alike).

limited by incentive and implementation problems, and may require intervention from other disciplines to create a hospitable market and regulatory climate.[68] Furthermore, some privacy-enhancing AI applications may merely remove obstacles to the use of other privacy-eroding activities, resulting in little or no gain in overall privacy protection.[69] Maintaining appropriate privacy guarantees is a difficult and constantly evolving task that will require a multi-pronged approach. This Part outlines a few non-technical considerations that may help accomplish this goal.

A. Private Reinforcement

In order to promote uptake of AI PETs, consumers must be educated about the current risks to their digital privacy and instructed in basic digital literacy. Without this foundation, there will be no demand for privacy-enhancing technology, nor will PETs be effectively used. Consumers will need to be aware of differential privacy and federated learning so that they will know to adopt platforms with these privacy protections, or demand that the platforms they use adopt them. Using personal AI guardians will also likely require a considerable amount of digital literacy, from setting up and administering the guardian to interpreting any feedback it provides. In particular, users of any AI systems with overrides should be educated about when they should expect to take matters into their own hands. Education of any kind will necessarily require teaching others to value data privacy so that the skills learned will be put to use. Some scholars have even suggested that protecting one's own personal privacy should be treated as a moral duty.[70]

Businesses, in turn, can be expected to leverage AI PETs to some degree as a means of protecting their reputations and the ability to compete in the marketplace. Indeed, this is already taking place; Apple and Microsoft have been heavily involved in the development of differential privacy, and Google is responsible for developing federated learning.[71]

[68] *See* GASSER, *supra* note 5; *cf.* Urs Gasser, *AI and the Law: Setting the Stage*, MEDIUM (June 26, 2017), https://medium.com/berkman-klein-center/ai-and-the-law-setting-the-stage-48516fda1b11 [https://perma.cc/BX3B-X2K5] ("At a fundamental level, a governance approach to AI-based technologies embraces and activates a variety of modes of regulation, including technology, social norms, markets and law, and combines these instruments with a blended governance framework. . . . From this 'blended governance' perspective, the main challenge is to identify and activate the most efficient, effective, and legitimate modalities for any given issue, and to successfully orchestrate the interplay among them.").

[69] As new technologies alleviate privacy risks at one stage in the data lifecycle — consider, for example, federated learning's ability to avoid the collection and storage of individualized data — businesses may be tempted to incorporate them into increasingly intrusive applications that would not otherwise have been pursued.

[70] *See* Anita L. Allen, *Protecting One's Own Privacy in a Big Data Economy*, 130 HARV. L. REV. FORUM 71, 72 (2016).

[71] *See supra* Sections II.A and II.B.

The degree to which further progress is prioritized will depend on the degree to which consumers demand it. Technology companies (and the trade associations they participate in) will also develop privacy solutions to the extent that doing so benefits their own innovation agenda. To them, self-regulation is generally preferable to legal regulation — which again is a function of societal demand — and some PETs will likely unblock new innovations that otherwise would have been too risky to develop.

Data brokers also play a large role in the digital privacy ecosystem, and will entertain a difficult set of choices with the advent of differential privacy (and federated learning, which could drastically shrink their businesses by cutting them out as middlemen). They will "face a choice between roughly three alternatives: sticking with the old habit of de-identification and hoping for the best; turning to emerging technologies like differential privacy that involve some trade-offs in utility and convenience; and using legal agreements to limit the flow and use of sensitive data."[72] The path data brokers take could be heavily shaped by regulation, especially if the law is used in support of newly proposed models such as information fiduciaries, discussed in the next Section.

B. Legal Reinforcement

Of special importance is how AI PETs will be enabled through law. Professor Urs Gasser has suggested that there are three main ways in which law responds to technological change: subsumption, gradual innovation, and paradigm shifts.[73] While certain entities may be incentivized to develop AI PETs due to the straightforward application of existing laws — for example, in order to obtain intellectual property rights over a new AI-related invention, or to minimize the risk of tort or statutory liability — AI is sufficiently distinct from other technologies. Many believe it can be adequately addressed only through gradual innovation or paradigm shifts. This Section will focus on each in turn.

1. Gradual Innovation

A quintessential example of gradual innovation is the development of new laws that largely resemble previous statutory models. AI PETs could be incentivized through new laws providing subsidies or safe harbors to entities that utilize state-of-the-art algorithmic data anonymization techniques, such as differential privacy and federated learning.

[72] Arvind Narayanan & Edward W. Felten, *No Silver Bullet: De-Identification Still Doesn't Work* (July 9, 2014) (unpublished manuscript), http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf [https://perma.cc/6K96-4BJY].

[73] *See* Urs Gasser, *Recoding Privacy Law: Reflections on the Future Relationship Among Law, Technology, and Privacy*, 130 HARV. L. REV. FORUM 61, 64 (2016).

Professor Matthew Scherer has extended this idea, suggesting that a regulatory body should certify uses of AI — whether privacy enhancing or not — according to minimum standards of safety and reliability, and that certified AI programs that cause legally cognizable harm could receive better legal presumptions.[74] Because of AI's somewhat unpredictable nature and its many socially beneficial uses, incentivizing responsible development through rewards rather than penalties may be most desirable.

However, there are some areas where penalties do make sense. The FTC might issue new interpretations as to what constitutes a deceptive or unfair business practice, for example, that cover certain forms of AI-based psychological manipulation.[75] Legislatures may also pass new privacy statutes to provide special protections for certain sensitive data, as they have done in the past in the health and financial sectors.[76] If companies apply differential privacy and federated learning more broadly, consumers may become willing to part with new forms of sensitive (or granular) data that previously would have been unthinkable. They will need recourse if that data is not handled responsibly.

2. Paradigm Shifts

A paradigm shift is a "deeply layered law reform" that fundamentally alters the role of law in a given field.[77] Generic examples of proposed paradigm shifts in laws governing technology include calls for digital

[74] In the context of tort claims, systems certified by the regulatory agency would enjoy a limited tort liability standard (actual negligence) rather than strict liability. *See* Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J.L. & TECH. 354, 394 (2016).

[75] While the FTC has not yet passed any regulations in this area, former Chairwoman Edith Ramirez and Acting Chairwoman Maureen Ohlhausen have both acknowledged potential harms to consumers caused by AI. *See* Natasha Lomas, *The FTC Warns Internet of Things Businesses to Bake In Privacy and Security*, TECHCRUNCH (Jan. 8, 2015), https://techcrunch.com/2015/01/08/ftc-iot-privacy-warning/ [https://perma.cc/FN5M-VWLZ] (quoting Edith Ramirez); David McCabe, *FTC May Take a Deeper Look at Artificial Intelligence*, AXIOS (Aug. 22, 2017), https://www.axios.com/ftc-may-take-a-deeper-look-at-artificial-intelligence-2475755741.html [https://perma.cc/ET63-4T3S] (quoting Maureen Ohlhausen).

[76] *See, e.g.*, Financial Services Modernization Act (Gramm-Leach-Bliley Act), 15 U.S.C. §§ 6801–6827 (2012) (regarding financial information); Fair Credit Reporting Act (FCRA), 15 U.S.C. § 1681 (2012) (regarding credit information); Health Insurance Portability and Accountability Act (HIPAA), 42 U.S.C. § 1301 (2012) (regarding health information). It is unclear whether further legislation will be passed anytime soon. While the Obama administration tried unsuccessfully in recent years to promulgate a Consumer Privacy Bill of Rights and encourage related legislation, the Trump administration has not signaled any intent to continue that effort, and even removed a newly developed digital privacy report from the White House website shortly after the inauguration. Kate Kaye, *New Privacy Report Already Removed from White House Site*, ADVERTISINGAGE (Jan. 20, 2017), http://adage.com/article/privacy-and-regulation/privacy-report-removed-white-house-site/307632/ [https://perma.cc/4QLX-ZSJA].

[77] *See* Gasser, *supra* note 73, at 64–65.

dispute resolution systems[78] and recognition of international regimes to resolve cross-border digital disputes.[79] In the context of AI PETs, there are three other foreseeable paradigm shifts. First, using AI to refine the definition of privacy and privacy harms could have significant implications for privacy law, including clarification from the Supreme Court of what constitutes an injury-in-fact for Article III standing purposes. As another example, interpreting privacy as contextual integrity may lead to a more nuanced understanding of consent with respect to information disclosure, thus facilitating the development of informational privacy torts and revenge porn legislation.[80] Second, not unlike Scherer's call for an AI regulatory body, a government oversight entity could employ AI auditors to analyze other AI systems for fairness and accuracy. The systems monitored could include those belonging to the government, those belonging to businesses, or a subset of systems deemed especially crucial to society — something akin to "common carrier" systems. Monitoring could also be carried out by trusted private entities or a quasi-government entity; the process would likely consume substantial resources because constantly evolving AI systems would need to be reexamined on an ongoing basis.

Third and finally, Congress could create legislation regulating information fiduciaries. The idea of an information fiduciary was originally developed by Professors Jack Balkin and Jonathan Zittrain, who use the term to describe an information trader with a duty to act in a trustworthy manner with respect to that information.[81] Information fiduciar-

[78] *Id.* at 64 (referencing idea of a 'cyber court').

[79] *See* NORMAN SOLOVAY & CYNTHIA K. REED, THE INTERNET AND DISPUTE RESOLUTION §§ 9.01–9.02 (2003) (contemplating international cooperation through organizations such as UNCITRAL to effectuate online dispute resolution ("ODR") procedures and resolve choice-of-law and jurisdictional conflicts); Pietro Ortolani, *Self-Enforcing Online Dispute Resolution: Lessons from Bitcoin*, 36 OXFORD J. LEGAL STUD. 595, 595–96, 598–602 (2016) (proposing use of the Bitcoin system, an algorithm underlying many smart contracts, to refine cross-border ODR). *See generally* Laurence R. Helfer, *International Dispute Settlement at the Trademark-Doman Name Interface*, 29 PEPP. L. REV. 87 (2001) (describing ICANN's Uniform Domain Name Dispute Resolution Policy, an unconventional system used to resolve conflicts between domain name and trademark owners).

[80] *Cf.* A. Michael Froomkin, *The Death of Privacy?*, 52 STAN. L. REV. 1461, 1535–37 (2000); Danielle Keats Citron & Mary Anne Franks, *Criminalizing Revenge Porn*, 49 WAKE FOREST L. REV. 345, 348 (2014) (citing Nissenbaum's theory of contextual integrity to argue in the context of revenge porn that "[i]ndividual and societal expectations of privacy are tailored to specific circumstances. The nonconsensual sharing of an individual's intimate photos should be no different; consent within a trusted relationship does not equal consent outside of that relationship").

[81] Jack M. Balkin, *Information Fiduciaries in the Digital Age*, BALKINIZATION (Mar. 5, 2014), https://balkin.blogspot.com/2014/03/information-fiduciaries-in-digital-age.html [https://perma.cc/9GHQ-L2Q7]; *see also* Jonathan Zittrain, *Facebook Could Decide an Election Without Anyone Ever Finding Out: The Scary Future of Digital Gerrymandering — and How to Prevent It*, NEW REPUBLIC (June 1, 2014), https://newrepublic.com/article/117878/information-fiduciary-solution-facebook-digital-gerrymandering (last visited Dec. 20, 2017).

ies would have obligations similar to those of a doctor, lawyer, or accountant to keep certain information confidential and not to use clients' information against them.[82] This system creates trust where there otherwise would not be,[83] allowing parties to safely exchange sensitive information. In short, it forces fiduciaries to behave in a manner that better aligns with social expectations of how sensitive information should be handled.[84]

The duties of confidentiality and competence commonly associated with fiduciaries are usually interpreted under a reasonableness standard in light of the surrounding circumstances.[85] Consequently, the information practices and security guarantees offered by information fiduciaries will depend on the state of available privacy technologies, including AI PETs. Differential privacy and federated learning, as they are developed further, should dictate that information fiduciaries minimize their collection and storage of raw user data, and provide approved third parties with strongly anonymized or artificial data whenever possible. Because differential privacy methods can support query-based access to data more robustly than wholesale dataset access,[86] there should also be a presumption in favor of providing third parties only query-based access. Lastly, reasonableness standards would accommodate input from AI that elucidates societal privacy expectations under a contextual integrity theory of privacy. AI could therefore help define the boundaries of the fiduciary duties themselves, or in other words, what it means to keep an end user's trust.

IV. CONCLUSION

While AI is responsible for many breakthroughs in modern technology, the widespread use of AI to make decisions about nearly every aspect of our lives has justifiably sparked controversy arising from the technology's many risks and limitations. Its impact on digital privacy is particularly concerning, as AI has led to ubiquitous data gathering, reidentification issues, and a lack of algorithmic accountability. This Note suggests that AI can also help *alleviate* many digital privacy challenges, however, and is not simply a privacy menace. The new techniques embodied in differential privacy and federated learning minimize the amount of sensitive information that is collected, stored, and shared with

[82] *See* Balkin, *supra* note 81.

[83] *See id.*

[84] Jack M. Balkin and Jonathan Zittrain, *A Grand Bargain to Make Tech Companies Trustworthy*, THE ATLANTIC (Oct. 3, 2016), https://www.theatlantic.com/technology/archive/2016/10/information-fiduciary/502346/ [https://perma.cc/4NTG-TM2M] ("[W]e have to trust online services, but we have no real guarantees that they will not abuse our trust.").

[85] *See, e.g.*, MODEL RULES OF PROF'L CONDUCT r. 1.6 cmt. 18–19 (AM. BAR ASS'N 2017).

[86] Rubenstein & Hartzog, *supra* note 24, at 743.

others, while enabling companies to continue learning valuable insights from their users' activities. AI auditors and guardians can represent the interests of consumers through collectivization and incentive correction, and can monitor the likelihood of re-identification or discriminatory outcomes in other systems. Finally, AI might help us define privacy itself, a task that has often proved elusive and undermined the ability to enact more meaningful privacy protections. This is just a survey of possibilities as they currently exist; they are not necessarily compatible with one another, and each have different strengths and weaknesses. Successful implementation of these techniques will also depend on coordinating legislation and private action to ensure they realize their full potential. With the right guidance, AI still stands to change our lives for the better.