SPEAKER SERIES 2018

LAW AND NEUROSCIENCE

FORDHAM UNIVERSITY

THE SCHOOL OF LAW

March 20, 2018 Class 9

Darby J. Green, Esq.
Commercial Director
Litigation & Bankruptcy
Bloomberg Law

BRAVE NEW WORLD: THE GROWTH OF AI AND ANALYTICS IN LAW

Reading Materials

1. Biographical Information on Darby J. Green, Esq.
2. Webinar*: What Lawyers Think About A.I. and Why They Are Wrong*
3. *Neuroscience-Inspired Artificial Intelligence*
4. *Machine Learning and Law*
5. *Digital Direction for the Analog Attorney—Data Protection, E-Discovery, and the Ethics of Technological Competence in Today's World of Tomorrow*
6. *AI Integration With Blockchain*

BIOGRAPHICAL INFORMATION ON DARBY J. GREEN, ESQ.

Darby Green is the Commercial Director for Litigation and Bankruptcy at Bloomberg Law. In this role, she is responsible for product development and go-to-market strategies focused on the business intelligence and legal research needs of litigators. Ms. Green is the Chair of the Bloomberg Law Litigation Innovation Board, a group of twenty top litigators who provide input and consultation on product offerings. She also spearheaded the creation and launch of Bloomberg Law Litigation Analytics, a tool that enables users to search millions of legal data points by company, law firm, or judge; and Points of Law, a solution that applies machine learning to court opinions in order to highlight language critical to a court's holding and link this language to governing statements of law and relevant on-point case law. Ms. Green has been working on Bloomberg Law since 2009, prior to which she practiced as a commercial litigator in New York. She has an A.B. from Dartmouth College and a J.D. from Vanderbilt University Law School.

WEBINAR:
WHAT LAWYERS THINK ABOUT A.I. AND WHY THEY ARE WRONG

Breaking Media 11.29.mp4

Here also is a link to the original description of the webinar:
https://abovethelaw.com/2017/11/free-webinar-what-lawyers-think-about-ai-and-why-theyre-wrong/

Neuron
Review

Neuroscience-Inspired Artificial Intelligence

Demis Hassabis,[1,2,*] Dharshan Kumaran,[1,3] Christopher Summerfield,[1,4] and Matthew Botvinick[1,2]
[1]DeepMind, 5 New Street Square, London, UK
[2]Gatsby Computational Neuroscience Unit, 25 Howland Street, London, UK
[3]Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London, UK
[4]Department of Experimental Psychology, University of Oxford, Oxford, UK
*Correspondence: dhcontact@google.com
http://dx.doi.org/10.1016/j.neuron.2017.06.011

The fields of neuroscience and artificial intelligence (AI) have a long and intertwined history. In more recent times, however, communication and collaboration between the two fields has become less commonplace. In this article, we argue that better understanding biological brains could play a vital role in building intelligent machines. We survey historical interactions between the AI and neuroscience fields and emphasize current advances in AI that have been inspired by the study of neural computation in humans and other animals. We conclude by highlighting shared themes that may be key for advancing future research in both fields.

In recent years, rapid progress has been made in the related fields of neuroscience and artificial intelligence (AI). At the dawn of the computer age, work on AI was inextricably intertwined with neuroscience and psychology, and many of the early pioneers straddled both fields, with collaborations between these disciplines proving highly productive (Churchland and Sejnowski, 1988; Hebb, 1949; Hinton et al., 1986; Hopfield, 1982; McCulloch and Pitts, 1943; Turing, 1950). However, more recently, the interaction has become much less commonplace, as both subjects have grown enormously in complexity and disciplinary boundaries have solidified. In this review, we argue for the critical and ongoing importance of neuroscience in generating ideas that will accelerate and guide AI research (see Hassabis commentary in Brooks et al., 2012).

We begin with the premise that building human-level general AI (or "Turing-powerful" intelligent systems; Turing, 1936) is a daunting task, because the search space of possible solutions is vast and likely only very sparsely populated. We argue that this therefore underscores the utility of scrutinizing the inner workings of the human brain— the only existing proof that such an intelligence is even possible. Studying animal cognition and its neural implementation also has a vital role to play, as it can provide a window into various important aspects of higher-level general intelligence.

The benefits to developing AI of closely examining biological intelligence are two-fold. First, neuroscience provides a rich source of *inspiration* for new types of algorithms and architectures, independent of and complementary to the mathematical and logic-based methods and ideas that have largely dominated traditional approaches to AI. For example, were a new facet of biological computation found to be critical to supporting a cognitive function, then we would consider it an excellent candidate for incorporation into artificial systems. Second, neuroscience can provide *validation* of AI techniques that already exist. If a known algorithm is subsequently found to be implemented in the brain, then that is strong support for its plausibility as an integral component of an overall general intelligence system. Such clues can be critical to a long-term research program when determining where to allocate resources most productively. For example, if an algorithm is not quite attaining the level of performance required or expected, but we observe it is core to the functioning of the brain, then we can surmise that redoubled engineering efforts geared to making it work in artificial systems are likely to pay off.

Of course from a practical standpoint of building an AI system, we need not slavishly enforce adherence to biological plausibility. From an engineering perspective, what works is ultimately all that matters. For our purposes then, biological plausibility is a guide, not a strict requirement. What we are interested in is a systems neuroscience-level understanding of the brain, namely the algorithms, architectures, functions, and representations it utilizes. This roughly corresponds to the top two levels of the three levels of analysis that Marr famously stated are required to understand any complex biological system (Marr and Poggio, 1976): the goals of the system (the computational level) and the process and computations that realize this goal (the algorithmic level). The precise mechanisms by which this is physically realized in a biological substrate are less relevant here (the implementation level). Note this is where our approach to neuroscience-inspired AI differs from other initiatives, such as the Blue Brain Project (Markram, 2006) or the field of neuromorphic computing systems (Esser et al., 2016), which attempt to closely mimic or directly reverse engineer the specifics of neural circuits (albeit with different goals in mind). By focusing on the computational and algorithmic levels, we gain transferrable insights into general mechanisms of brain function, while leaving room to accommodate the distinctive opportunities and challenges that arise when building intelligent machines *in silico*.

The following sections unpack these points by considering the past, present, and future of the AI-neuroscience interface. Before beginning, we offer a clarification. Throughout this article, we employ the terms "neuroscience" and "AI." We use these terms in the widest possible sense. When we say neuroscience, we mean to include all fields that are involved with the study of the brain, the behaviors that it generates, and the mechanisms by which it does so, including cognitive neuroscience, systems neuroscience and psychology. When we say AI, we mean work

in machine learning, statistics, and AI research that aims to build intelligent machines (Legg and Hutter, 2007).

We begin by considering the origins of two fields that are pivotal for current AI research, deep learning and reinforcement learning, both of which took root in ideas from neuroscience. We then turn to the current state of play in AI research, noting many cases where inspiration has been drawn (sometimes without explicit acknowledgment) from concepts and findings in neuroscience. In this section, we particularly emphasize instances where we have combined deep learning with other approaches from across machine learning, such as reinforcement learning (Mnih et al., 2015), Monte Carlo tree search (Silver et al., 2016), or techniques involving an external content-addressable memory (Graves et al., 2016). Next, we consider the potential for neuroscience to support future AI research, looking at both the most likely research challenges and some emerging neuroscience-inspired AI techniques. While our main focus will be on the potential for neuroscience to benefit AI, our final section will briefly consider ways in which AI may be helpful to neuroscience and the broader potential for synergistic interactions between these two fields.

The Past

Deep Learning

As detailed in a number of recent reviews, AI has been revolutionized over the past few years by dramatic advances in neural network, or "deep learning," methods (LeCun et al., 2015; Schmidhuber, 2014). As the moniker "neural network" might suggest, the origins of these AI methods lie directly in neuroscience. In the 1940s, investigations of neural computation began with the construction of artificial neural networks that could compute logical functions (McCulloch and Pitts, 1943). Not long after, others proposed mechanisms by which networks of neurons might learn incrementally via supervisory feedback (Rosenblatt, 1958) or efficiently encode environmental statistics in an unsupervised fashion (Hebb, 1949). These mechanisms opened up the field of artificial neural network research, and they continue to provide the foundation for contemporary research on deep learning (Schmidhuber, 2014).

Not long after this pioneering work, the development of the backpropagation algorithm allowed learning to occur in networks composed of multiple layers (Rumelhart et al., 1985; Werbos, 1974). Notably, the implications of this method for understanding intelligence, including AI, were first appreciated by a group of neuroscientists and cognitive scientists, working under the banner of parallel distributed processing (PDP) (Rumelhart et al., 1986). At the time, most AI research was focused on building logical processing systems based on serial computation, an approach inspired in part by the notion that human intelligence involves manipulation of symbolic representations (Haugeland, 1985). However, there was a growing sense in some quarters that purely symbolic approaches might be too brittle and inflexible to solve complex real-world problems of the kind that humans routinely handle. Instead, a growing foundation of knowledge about the brain seemed to point in a very different direction, highlighting the role of stochastic and highly parallelized information processing. Building on this, the PDP movement proposed that human cognition and behavior emerge from dynamic, distributed interactions within networks of simple neuron-like processing units, interactions tuned by learning procedures that adjust system parameters in order to minimize error or maximize reward.

Although the PDP approach was at first applied to relatively small-scale problems, it showed striking success in accounting for a wide range of human behaviors (Hinton et al., 1986). Along the way, PDP research introduced a diverse collection of ideas that have had a sustained influence on AI research. For example, current machine translation research exploits the notion that words and sentences can be represented in a distributed fashion (i.e., as vectors) (LeCun et al., 2015), a principle that was already ingrained in early PDP-inspired models of sentence processing (St. John and McClelland, 1990). Building on the PDP movement's appeal to biological computation, current state-of-the-art convolutional neural networks (CNNs) incorporate several canonical hallmarks of neural computation, including nonlinear transduction, divisive normalization, and maximum-based pooling of inputs (Yamins and DiCarlo, 2016). These operations were directly inspired by single-cell recordings from the mammalian visual cortex that revealed how visual input is filtered and pooled in simple and complex cells in area V1 (Hubel and Wiesel, 1959). Moreover, current network architectures replicate the hierarchical organization of mammalian cortical systems, with both convergent and divergent information flow in successive, nested processing layers (Krizhevsky et al., 2012; LeCun et al., 1989; Riesenhuber and Poggio, 1999; Serre et al., 2007), following ideas first advanced in early neural network models of visual processing (Fukushima, 1980). In both biological and artificial systems, successive non-linear computations transform raw visual input into an increasingly complex set of features, permitting object recognition that is invariant to transformations of pose, illumination, or scale.

As the field of deep learning evolved out of PDP research into a core area within AI, it was bolstered by new ideas, such as the development of deep belief networks (Hinton et al., 2006) and the introduction of large datasets inspired by research on human language (Deng et al., 2009). During this period, it continued to draw key ideas from neuroscience. For example, biological considerations informed the development of successful regularization schemes that support generalization beyond training data. One such scheme, in which only a subset of units participate in the processing of a given training example ("dropout"), was motivated by the stochasticity that is inherent in biological systems populated by neurons that fire with Poisson-like statistics (Hinton et al., 2012). Here and elsewhere, neuroscience has provided initial guidance toward architectural and algorithmic constraints that lead to successful neural network applications for AI.

Reinforcement Learning

Alongside its important role in the development of deep learning, neuroscience was also instrumental in erecting a second pillar of contemporary AI, stimulating the emergence of the field of reinforcement learning (RL). RL methods address the problem of how to maximize future reward by mapping states in the environment to actions and are among the most widely used tools in AI research (Sutton and Barto, 1998). Although it is not widely appreciated among AI researchers, RL methods were originally

inspired by research into animal learning. In particular, the development of temporal-difference (TD) methods, a critical component of many RL models, was inextricably intertwined with research into animal behavior in conditioning experiments. TD methods are real-time models that learn from differences between temporally successive predictions, rather than having to wait until the actual reward is delivered. Of particular relevance was an effect called second-order conditioning, where affective significance is conferred on a conditioned stimulus (CS) through association with another CS rather than directly via association with the unconditioned stimulus (Sutton and Barto, 1981). TD learning provides a natural explanation for second-order conditioning and indeed has gone on to explain a much wider range of findings from neuroscience, as we discuss below.

Here, as in the case of deep learning, investigations initially inspired by observations from neuroscience led to further developments that have strongly shaped the direction of AI research. From their neuroscience-informed origins, TD methods and related techniques have gone on to supply the core technology for recent advances in AI, ranging from robotic control (Hafner and Riedmiller, 2011) to expert play in backgammon (Tesauro, 1995) and Go (Silver et al., 2016).

The Present

Reading the contemporary AI literature, one gains the impression that the earlier engagement with neuroscience has diminished. However, if one scratches the surface, one can uncover many cases in which recent developments have been inspired and guided by neuroscientific considerations. Here, we look at four specific examples.

Attention

The brain does not learn by implementing a single, global optimization principle within a uniform and undifferentiated neural network (Marblestone et al., 2016). Rather, biological brains are modular, with distinct but interacting subsystems underpinning key functions such as memory, language, and cognitive control (Anderson et al., 2004; Shallice, 1988). This insight from neuroscience has been imported, often in an unspoken way, into many areas of current AI.

One illustrative example is recent AI work on attention. Up until quite lately, most CNN models worked directly on entire images or video frames, with equal priority given to all image pixels at the earliest stage of processing. The primate visual system works differently. Rather than processing all input in parallel, visual attention shifts strategically among locations and objects, centering processing resources and representational coordinates on a series of regions in turn (Koch and Ullman, 1985; Moore and Zirnsak, 2017; Posner and Petersen, 1990). Detailed neurocomputational models have shown how this piecemeal approach benefits behavior, by prioritizing and isolating the information that is relevant at any given moment (Olshausen et al., 1993; Salinas and Abbott, 1997). As such, attentional mechanisms have been a source of inspiration for AI architectures that take "glimpses" of the input image at each step, update internal state representations, and then select the next location to sample (Larochelle and Hinton, 2010; Mnih et al., 2014) (Figure 1A). One such network was able to use this selective attentional mechanism to ignore irrelevant objects in a scene, allowing it to perform well in challenging object classification tasks in the presence of clutter (Mnih et al., 2014). Further, the attentional mechanism allowed the computational cost (e.g., number of network parameters) to scale favorably with the size of the input image. Extensions of this approach were subsequently shown to produce impressive performance at difficult multi-object recognition tasks, outperforming conventional CNNs that process the entirety of the image, both in terms of accuracy and computational efficiency (Ba et al., 2015), as well as enhancing image-to-caption generation (Xu et al., 2015).

While attention is typically thought of as an orienting mechanism for perception, its "spotlight" can also be focused internally, toward the contents of memory. This idea, a recent focus in neuroscience studies (Summerfield et al., 2006), has also inspired work in AI. In some architectures, attentional mechanisms have been used to select information to be read out from the internal memory of the network. This has helped provide recent successes in machine translation (Bahdanau et al., 2014) and led to important advances on memory and reasoning tasks (Graves et al., 2016). These architectures offer a novel implementation of content-addressable retrieval, which was itself a concept originally introduced to AI from neuroscience (Hopfield, 1982).

One further area of AI where attention mechanisms have recently proven useful focuses on generative models, systems that learn to synthesize or "imagine" images (or other kinds of data) that mimic the structure of examples presented during training. Deep generative models (i.e., generative models implemented as multi-layered neural networks) have recently shown striking successes in producing synthetic outputs that capture the form and structure of real visual scenes via the incorporation of attention-like mechanisms (Hong et al., 2015; Reed et al., 2016). For example, in one state-of-the-art generative model known as DRAW, attention allows the system to build up an image incrementally, attending to one portion of a "mental canvas" at a time (Gregor et al., 2015).

Episodic Memory

A canonical theme in neuroscience is that that intelligent behavior relies on multiple memory systems (Tulving, 1985). These will include not only reinforcement-based mechanisms, which allow the value of stimuli and actions to be learned incrementally and through repeated experience, but also instance-based mechanisms, which allow experiences to be encoded rapidly (in "one shot") in a content-addressable store (Gallistel and King, 2009). The latter form of memory, known as episodic memory (Tulving, 2002), is most often associated with circuits in the medial temporal lobe, prominently including the hippocampus (Squire et al., 2004).

One recent breakthrough in AI has been the successful integration of RL with deep learning (Mnih et al., 2015; Silver et al., 2016). For example, the deep Q-network (DQN) exhibits expert play on Atari 2600 video games by learning to transform a vector of image pixels into a policy for selecting actions (e.g., joystick movements). One key ingredient in DQN is "experience replay," whereby the network stores a subset of the training data in an instance-based way, and then "replays" it offline, learning anew from successes or failures that occurred in the past. Experience replay is critical to maximizing data efficiency, avoids the

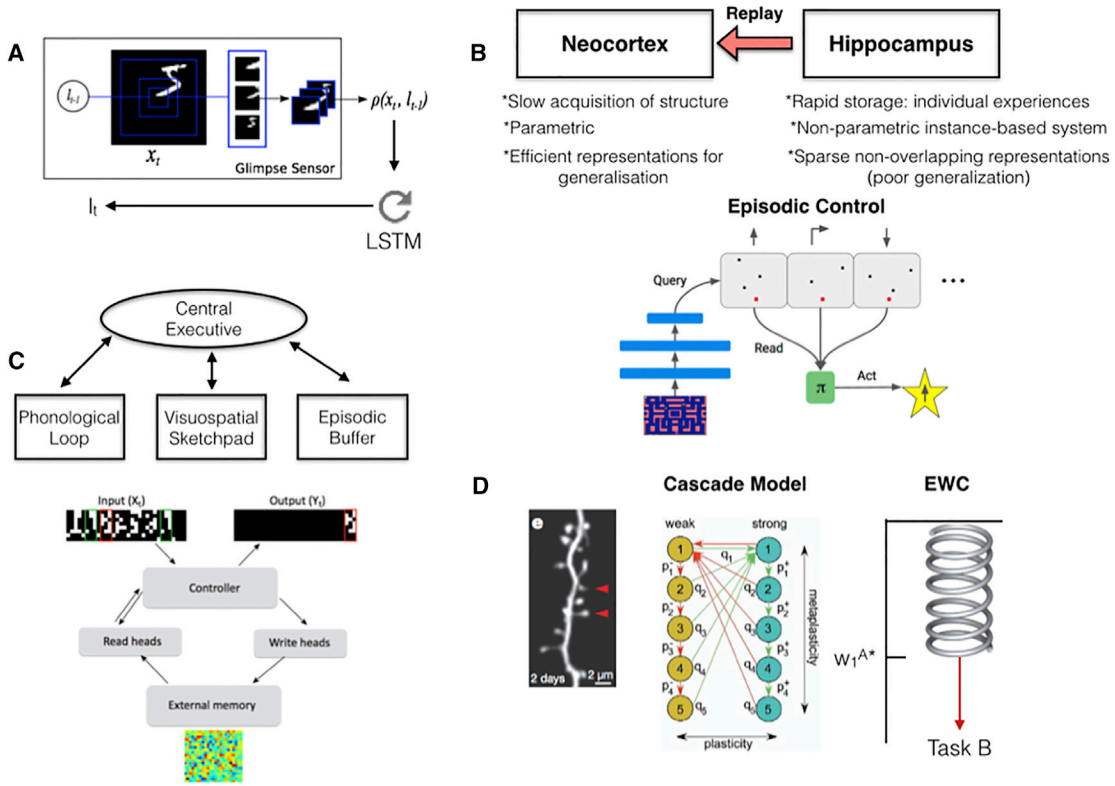

Figure 1. Parallels between AI Systems and Neural Models of Behavior

(A) Attention. Schematic of recurrent attention model (Mnih et al., 2014). Given an input image (x_t) and foveal location (l_{t-1}), the glimpse sensor extracts a multi-resolution "retinal" representation ($\rho(x_t, l_{t-1})$). This is the input to a glimpse network, which produces a representation that is passed to the LSTM core, which defines the next location to attend to (l_t) (and classification decision).

(B) Schematic of complementary learning systems and episodic control. Top: non-parametric fast learning hippocampal system and parametric slow-learning neocortical system (i.e., parametric: a fixed number of parameters; non-parametric: the number of parameters can grow with the amount of data). Hippocampus/instance-based system supports rapid behavioral adjustment (i.e., episodic control; Blundell et al., 2016) and experience replay, which supports interleaved training (i.e., on random subsets of experiences) of deep neural network (Mnih et al., 2015) or neocortex. Bottom: episodic control (from Blundell et al., 2016). Game states (Atari shown) are stored within buffers (one for each possible action) together with the highest (discounted) return experienced from that state (i.e., Q-value). When experiencing a new state, the policy (π) is determined by averaging the Q-value across the k nearest neighbors in each action buffer and selecting the action with the highest expected return.

(C) Illustration of parallels between macroscopic organization of models of working memory and the differentiable neural computer (Graves et al., 2016) (or Neural Turing Machine). The network controller (typically recurrent) is analogous to the central executive (typically viewed to be instantiated in the prefrontal cortex) and attends/reads/writes to an external memory matrix (phonological loo /sketchpad in working memory model). Architecture is shown performing copy task.

(D) Illustration of parallel between neurobiological models of synaptic consolidation and the elastic weight consolidation (EWC) algorithm. Left: two-photon structural imaging data showing learning-related increase in size of dendrites (each corresponding approximately to a single excitatory synapse) that persists for months (from Yang et al., 2009). Middle: schematic of Cascade model of synaptic consolidation (adapted with permission from Fusi et al., 2005). Binary synapses transition between metaplastic states which are more/less plastic (least plastic states at bottom of diagram), as a function of prior potentiation/depression events. Right panel: schematic of elastic weight consolidation (EWC) algorithm. After training on the first task (A), network parameters are optimized for good performance: single weight (w_1^{A*} illustrated). EWC implements a constraint analogous to a spring that anchors weights to the previously found solution (i.e., for task A), when training on a new task (e.g., task B), with the stiffness of the spring *proportional to the importance* of that parameter for task A performance (Kirkpatrick et al., 2017).

destabilizing effects of learning from consecutive correlated experiences, and allows the network to learn a viable value function even in complex, highly structured sequential environments such as video games.

Critically, experience replay was directly inspired by theories that seek to understand how the multiple memory systems in the mammalian brain might interact. According to a prominent view, animal learning is supported by parallel or "complementary" learning systems in the hippocampus and neocortex (Kumaran et al., 2016; McClelland et al., 1995). The hippocampus acts to encode novel information after a single exposure (one-shot learning), but this information is gradually consolidated to the neocortex in sleep or resting periods that are interleaved with periods of activity. This consolidation is accompanied by replay in the hippocampus and neocortex, which is observed as a reinstatement of the structured patterns of neural activity that accompanied the learning event (O'Neill et al., 2010; Skaggs and McNaughton, 1996) (Figure 1B). This theory was originally proposed as a solution to the well-known problem that in conventional neural networks, correlated exposure to sequential task settings leads to mutual interference among policies, resulting in catastrophic forgetting of one task as a new one is

learned. The replay buffer in DQN might thus be thought of as a very primitive hippocampus, permitting complementary learning *in silico* much as is proposed for biological brains. Later work showed that the benefits of experience replay in DQN are enhanced when replay of highly rewarding events is prioritized (Schaul et al., 2015), just as hippocampal replay seems to favor events that lead to high levels of reinforcement (Singer and Frank, 2009).

Experiences stored in a memory buffer can not only be used to gradually adjust the parameters of a deep network toward an optimal policy, as in DQN, but can also support rapid behavioral change based on an individual experience. Indeed, theoretical neuroscience has argued for the potential benefits of *episodic control*, whereby rewarded action sequences can be internally re-enacted from a rapidly updateable memory store, implemented in the biological case in the hippocampus (Gershman and Daw, 2017). Further, normative accounts show that episodic control is particularly advantageous over other learning mechanisms when limited experience of the environment has been obtained (Lengyel and Dayan, 2007).

Recent AI research has drawn on these ideas to overcome the slow learning characteristics of deep RL networks, developing architectures that implement episodic control (Blundell et al., 2016). These networks store specific experiences (e.g., actions and reward outcomes associated with particular Atari game screens) and select new actions based on the similarity between the current situation input and the previous events stored in memory, taking the reward associated with those previous events into account (Figure 1B). As predicted from the initial, neuroscience-based work (Lengyel and Dayan, 2007), artificial agents employing episodic control show striking gains in performance over deep RL networks, particularly early on during learning (Blundell et al., 2016). Further, they are able to achieve success on tasks that depend heavily on one-shot learning, where typical deep RL architectures fail. Moreover, episodic-like memory systems more generally have shown considerable promise in allowing new concepts to be learned rapidly based on only a few examples (Vinyals et al., 2016). In the future, it will be interesting to harness the benefits of rapid episodic-like memory and more traditional incremental learning in architectures that incorporate both of these components within an interacting framework that mirrors the complementary learning systems in mammalian brain. We discuss these future perspectives below in more detail later, in "Imagination and planning."

Working Memory

Human intelligence is characterized by a remarkable ability to maintain and manipulate information within an active store, known as working memory, which is thought to be instantiated within the prefrontal cortex and interconnected areas (Goldman-Rakic, 1990). Classic cognitive theories suggest that this functionality depends on interactions between a central controller ("executive") and separate, domain-specific memory buffers (e.g., visuo-spatial sketchpad) (Baddeley, 2012). AI research has drawn inspiration from these models, by building architectures that explicitly maintain information over time. Historically, such efforts began with the introduction of recurrent neural network architectures displaying attractor dynamics and rich sequential behavior, work directly inspired by neuroscience (Elman, 1990; Hopfield and Tank, 1986; Jordan, 1997). This work enabled later, more detailed modeling of human working memory (Botvinick and Plaut, 2006; Durstewitz et al., 2000), but it also laid the foundation for further technical innovations that have proved pivotal in recent AI research. In particular, one can see close parallels between the learning dynamics in these early, neuroscience-inspired networks and those in long-short-term memory (LSTM) networks, which subsequently achieved state of the art performance across a variety of domains. LTSMs allow information to be gated into a fixed activity state and maintained until an appropriate output is required (Hochreiter and Schmidhuber, 1997). Variants of this type of network have shown some striking behaviors in challenging domains, such as learning to respond to queries about the latent state of variables after training on computer code (Zaremba and Sutskever, 2014).

In ordinary LSTM networks, the functions of sequence control and memory storage are closely intertwined. This contrasts with classic models of human working memory, which, as mentioned above, separate these two. This neuroscience-based schema has recently inspired more complex AI architectures where control and storage are supported by distinct modules (Graves et al., 2014, 2016; Weston et al., 2014). For example, the differential neural computer (DNC) involves a neural network controller that attends to and reads/writes from an external memory matrix (Graves et al., 2016). This externalization allows the network controller to learn from scratch (i.e., via end-to-end optimization) to perform a wide range of complex memory and reasoning tasks that currently elude LSTMs, such as finding the shortest path through a graph-like structure, such as a subway map, or manipulating blocks in a variant of the Tower of Hanoi task (Figure 1C). These types of problems were previously argued to depend exclusively on symbol processing and variable binding and therefore beyond the purview of neural networks (Fodor and Pylyshyn, 1988; Marcus, 1998). Of note, although both LSTMs and the DNC are described here in the context of working memory, they have the potential to maintain information over many thousands of training cycles and so may thus be suited to longer-term forms of memory, such as retaining and understanding the contents of a book.

Continual Learning

Intelligent agents must be able to learn and remember many different tasks that are encountered over multiple timescales. Both biological and artificial agents must thus have a capacity for continual learning, that is, an ability to master new tasks without forgetting how to perform prior tasks (Thrun and Mitchell, 1995). While animals appear relatively adept at continual learning, neural networks suffer from the problem of catastrophic forgetting (French, 1999; McClelland et al., 1995). This occurs as the network parameters shift toward the optimal state for performing the second of two successive tasks, overwriting the configuration that allowed them to perform the first. Given the importance of continual learning, this liability of neural networks remains a significant challenge for the development of AI.

In neuroscience, advanced neuroimaging techniques (e.g., two-photon imaging) now allow dynamic in vivo visualization of the structure and function of dendritic spines during learning, at the spatial scale of single synapses (Nishiyama and Yasuda, 2015). This approach can be used to study

neocortical plasticity during continual learning (Cichon and Gan, 2015; Hayashi-Takagi et al., 2015; Yang et al., 2009). There is emerging evidence for specialized mechanisms that protect knowledge about previous tasks from interference during learning on a new task. These include decreased synaptic lability (i.e., lower rates of plasticity) in a proportion of strengthened synapses, mediated by enlargements to dendritic spines that persist despite learning of other tasks (Cichon and Gan, 2015; Yang et al., 2009) (Figure 1D). These changes are associated with retention of task performance over several months, and indeed, if they are "erased" with synaptic optogenetics, this leads to forgetting of the task (Hayashi-Takagi et al., 2015). These empirical insights are consistent with theoretical models that suggest that memories can be protected from interference through synapses that transition between a cascade of states with different levels of plasticity (Fusi et al., 2005) (Figure 1D).

Together, these findings from neuroscience have inspired the development of AI algorithms that address the challenge of continual learning in deep networks by implementing of a form of "elastic" weight consolidation (EWC) (Kirkpatrick et al., 2017), which acts by slowing down learning in a subset of network weights identified as important to previous tasks, thereby anchoring these parameters to previously found solutions (Figure 1D). This allows multiple tasks to be learned without an increase in network capacity, with weights shared efficiently between tasks with related structure. In this way, the EWC algorithm allows deep RL networks to support continual learning at large scale.

The Future

In AI, the pace of recent research has been remarkable. Artificial systems now match human performance in challenging object recognition tasks (Krizhevsky et al., 2012) and outperform expert humans in dynamic, adversarial environments such as Atari video games (Mnih et al., 2015), the ancient board game of Go (Silver et al., 2016), and imperfect information games such as heads-up poker (Moravčík et al., 2017). Machines can autonomously generate synthetic natural images and simulations of human speech that are almost indistinguishable from their real-world counterparts (Lake et al., 2015; van den Oord et al., 2016), translate between multiple languages (Wu et al., 2016), and create "neural art" in the style of well-known painters (Gatys et al., 2015).

However, much work is still needed to bridge the gap between machine and human-level intelligence. In working toward closing this gap, we believe ideas from neuroscience will become increasingly indispensable. In neuroscience, the advent of new tools for brain imaging and genetic bioengineering have begun to offer a detailed characterization of the computations occurring in neural circuits, promising a revolution in our understanding of mammalian brain function (Deisseroth and Schnitzer, 2013). The relevance of neuroscience, both as a roadmap for the AI research agenda and as a source of computational tools is particularly salient in the following key areas.

Intuitive Understanding of the Physical World

Recent perspectives emphasize key ingredients of human intelligence that are already well developed in human infants but lacking in most AI systems (Gilmore et al., 2007; Gopnik and Schulz, 2004; Lake et al., 2016). Among these capabilities are knowledge of core concepts relating to the physical world, such as space, number, and objectness, which allow people to construct compositional mental models that can guide inference and prediction (Battaglia et al., 2013; Spelke and Kinzler, 2007).

AI research has begun to explore methods for addressing this challenge. For example, novel neural network architectures have been developed that interpret and reason about scenes in a humanlike way, by decomposing them into individual objects and their relations (Battaglia et al., 2016; Chang et al., 2016; Eslami et al., 2016) (Figures 2A and 2B). In some cases, this has resulted in human-level performance on challenging reasoning tasks (Santoro et al., 2017). In other work, deep RL has been used to capture the processes by which children gain commonsense understanding of the world through interactive experiments (Denil et al., 2016). Relatedly, deep generative models have been developed that are able to construct rich object models from raw sensory inputs (Higgins et al., 2016). These leverage constraints first identified in neuroscience, such as redundancy reduction (Barlow, 1959), which encourage the emergence of disentangled representations of independent factors such as shape and position (Figure 2C). Importantly, the latent representations learned by such generative models exhibit compositional properties, supporting flexible transfer to novel tasks (Eslami et al., 2016; Higgins et al., 2016; Rezende et al., 2016a). In the caption associated with Figure 2, we provide more detailed information about these networks.

Efficient Learning

Human cognition is distinguished by its ability to rapidly learn about new concepts from only a handful of examples, leveraging prior knowledge to enable flexible inductive inferences. In order to highlight this human ability as a challenge for AI, Lake and colleagues recently posed a "characters challenge" (Lake et al., 2016). Here, an observer must distinguish novel instances of an unfamiliar handwritten character from other, similar items after viewing only a single exemplar. Humans can perform this task well, but it is difficult for classical AI systems.

Encouragingly, recent AI algorithms have begun to make progress on tasks like the characters challenge, through both structured probabilistic models (Lake et al., 2015) and deep generative models based on the abovementioned DRAW model (Rezende et al., 2016b). Both classes of system can make inferences about a new concept despite a poverty of data and generate new samples from a single example concept (Figure 2D). Further, recent AI research has developed networks that "learn to learn," acquiring knowledge on new tasks by leveraging prior experience with related problems, to support one-shot concept learning (Santoro et al., 2016; Vinyals et al., 2016) and accelerating learning in RL tasks (Wang et al., 2016). Once again, this builds on concepts from neuroscience: learning to learn was first explored in studies of animal learning (Harlow, 1949), and has subsequently been studied in developmental psychology (Adolph, 2005; Kemp et al., 2010; Smith, 1995).

Transfer Learning

Humans also excel at generalizing or transferring generalized knowledge gained in one context to novel, previously unseen domains (Barnett and Ceci, 2002; Holyoak and Thagard, 1997). For

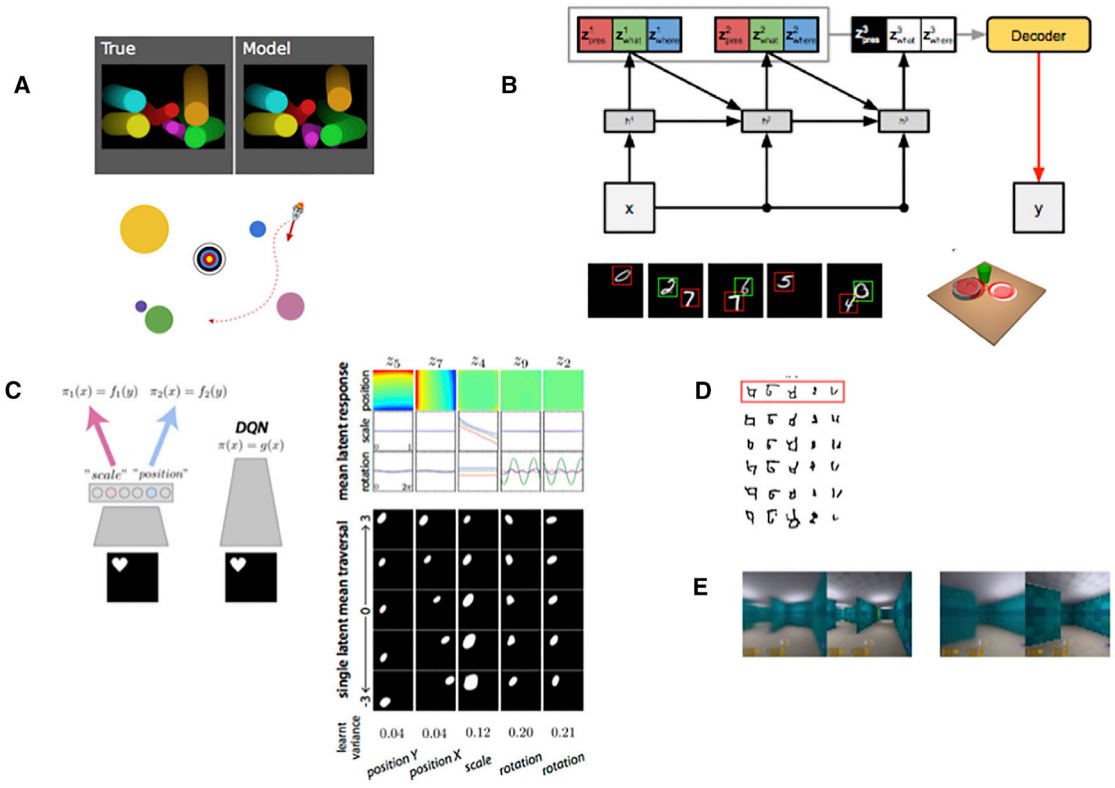

Figure 2. Examples of Recent AI Systems that Have Been Inspired by Neuroscience

(A) Intuitive physics knowledge. Illustration of the ability of the interaction network (Battaglia et al., 2016) to reason and make predictions about the physical interaction between objects in the bouncing ball problem (top) and spaceship problem (bottom: Hamrick et al., 2017). The network takes as input objects and their relations and accurately simulates their trajectories by modeling collisions, gravitational forces, etc., effectively acting as a learned physics engine.

(B) Scene understanding through structured generative models (Eslami et al., 2016). Top: iterative inference in a variational auto-encoder architecture. The recurrent network attends to one object at a time, infers its attributes, and performs the appropriate number of inference steps for each input image (x). Scenes are described in terms of groups of latent variables (Z) that specify presence/absence (z_{pres}), properties such as position (z_{where}), and shape (z_{what}). Inference network (black connections), and the generator network (red arrow), which produces reconstructed image (y). Bottom: illustration of iterative inference in multiple MNIST images (green indicates the first step and red the second step). Right: inference about the position/shape of multiple objects in realistic scene (note that inference is accurate, and hence it is difficult to distinguish inferred positions [red line] from ground truth). Latent representations in this network speed learning on downstream tasks (e.g., addition of MNIST digits) (not depicted; see Eslami et al., 2016).

(C) Unsupervised learning of core object properties (Higgins et al., 2016) is shown. Left: schematic illustrating learning of disentangled factors of sensory input by deep generative model (left: variational auto-encoder [VAE]), whose representations can speed learning on downstream tasks (Eslami et al., 2016), as compared to relatively entangled representation learned by typical deep network (e.g., DQN: right). Right panel illustrates latent representation of VAE; latent units coding for factors of variation, such as object position, rotation, and scale, are shown by effect of independently changing the activity of one latent unit. Such networks can learn intuitive concepts such as "objectness," being able to support zero-shot transfer (i.e., reasoning about position or scale of an unseen object with a novel shape; Higgins et al., 2016).

(D) One-shot generalization in deep sequential generative models (Rezende et al., 2016b) is shown. Deep generative models specify a causal process for generating the observed data using a hierarchy of latent variables, with attentional mechanisms supporting sequential inference. Illustrated are generated samples from the Rezende et al. model, conditioned on a single novel character from a held-out alphabet from the Omniglot dataset (Lake et al., 2015), demonstrating abilities that mirror human abilities to generalize from a single concept.

(E) Imagination of realistic environments in deep networks (Chiappa et al., 2017) is shown. Generated (left) and real (right) frames from procedural mazes (i.e., new maze layout on each episode) produced by an action-conditional recurrent network model ~150 and 200 frames after the last observed image, respectively.

example, a human who can drive a car, use a laptop computer, or chair a committee meeting is usually able act effectively when confronted with an unfamiliar vehicle, operating system, or social situation. Progress is being made in developing AI architectures capable of exhibiting strong generalization or transfer, for example by enabling zero-shot inferences about novel shapes outside the training distribution based on compositional representations (Higgins et al., 2016; Figure 2C). Others have shown that a new class of architecture, known as a progressive network, can leverage knowledge gained in one video game to learn rapidly in another, promising the sort of "far transfer" that is characteristic of human skill acquisition (Rusu et al., 2016a). Progressive networks have also been successfully employed to transfer knowledge for a simulated robotic environment to a real robot arm, massively reducing the training time required on the real world (Rusu et al., 2016b). Intriguingly, the proposed architecture bears some resemblance to a successful computational model of sequential task learning in humans (Collins and Koechlin, 2012; Donoso et al., 2014). In the neuroscience literature, one hallmark of transfer learning has been the ability to

reason relationally, and AI researchers have also begun to make progress in building deep networks that address problems of this nature, for example by solving visual analogies (Reed et al., 2015). More generally however, how humans or other animals achieve this sort of high-level transfer learning is unknown, and remains a relatively unexplored topic in neuroscience. New advances on this front could provide critical insights to spur AI research toward the goal of lifelong learning in agents, and we encourage neuroscientists to engage more deeply with this question.

At the level of neural coding, this kind of transfer of abstract structured knowledge may rely on the formation of conceptual representations that are invariant to the objects, individuals, or scene elements that populate a sensory domain but code instead for abstract, relational information among patterns of inputs (Doumas et al., 2008). However, we currently lack direct evidence for the existence of such codes in the mammalian brain. Nevertheless, one recent report made the very interesting claim that neural codes thought to be important in the representation of allocentric (map-like) spaces might be critical for abstract reasoning in more general domains (Constantinescu et al., 2016). In the mammalian entorhinal cortex, cells encode the geometry of allocentric space with a periodic "grid" code, with receptive fields that tile the local space in a hexagonal pattern (Rowland et al., 2016). Grid codes may be an excellent candidate for organizing conceptual knowledge, because they allow state spaces to be decomposed efficiently, in a way that could support discovery of subgoals and hierarchical planning (Stachenfeld et al., 2014). Using functional neuroimaging, the researchers provide evidence for the existence of such codes while humans performed an abstract categorization task, supporting the view that periodic encoding is a generalized hallmark of human knowledge organization (Constantinescu et al., 2016). However, much further work is required to substantiate this interesting claim.

Imagination and Planning

Despite their strong performance on goal-directed tasks, deep RL systems such as DQN operate mostly in a reactive way, learning the mapping from perceptual inputs to actions that maximize future value. This "model-free" RL is computationally inexpensive but suffers from two major drawbacks: it is relatively data inefficient, requiring large amounts of experience to derive accurate estimates, and it is inflexible, being insensitive to changes in the value of outcomes (Daw et al., 2005). By contrast, humans can more flexibly select actions based on forecasts of long-term future outcomes through simulation-based planning, which uses predictions generated from an internal model of the environment learned through experience (Daw et al., 2005; Dolan and Dayan, 2013; Tolman, 1948). Moreover, planning is not a uniquely human capacity. For example, when caching food, scrub jays consider the future conditions under which it is likely to be recovered (Raby et al., 2007), and rats use a "cognitive map" when navigating, allowing inductive inferences during wayfinding and facilitating one-shot learning behaviors in maze-like environments (Daw et al., 2005; Tolman, 1948). Of course, this point has not been lost on AI researchers; indeed, early planning algorithms such as Dyna (Sutton, 1991) were inspired by theories that emphasized the importance of "mental models" in generating hypothetical experiences useful for human learning (Craik, 1943). By now, a large volume of literature exists on AI planning techniques, including model-based RL methods, which seek to implement this forecast-based method of action selection. Furthermore, simulation-based planning, particularly Monte Carlo tree search (MCTS) methods, which use forward search to update a value function and/or policy (Browne et al., 2012), played a key role in recent work in which deep RL attained expert-level performance in the game of Go (Silver et al., 2016).

AI research on planning, however, has yet to capture some of the key characteristics that give human planning abilities their power. In particular, we suggest that a general solution to this problem will require understanding how rich internal models, which in practice will have to be approximate but sufficiently accurate to support planning, can be learned through experience, without strong priors being handcrafted into the network by the experimenter. We also argue that AI research will benefit from a close reading of the related literature on how humans imagine possible scenarios, envision the future, and carry out simulation-based planning, functions that depend on a common neural substrate in the hippocampus (Doll et al., 2015; Hassabis and Maguire, 2007, 2009; Schacter et al., 2012). Although imagination has an intrinsically subjective, unobservable quality, we have reason to believe that it has a conserved role in simulation-based planning across species (Hassabis and Maguire, 2009; Schacter et al., 2012). For example, when paused at a choice point, ripples of neural activity in the rat hippocampus resemble those observed during subsequent navigation of the available trajectories ("preplay"), as if the animal were "imagining" each possible alternative (Johnson and Redish, 2007; Ólafsdóttir et al., 2015; Pfeiffer and Foster, 2013). Further, recent work has suggested a similar process during non-spatial planning in humans (Doll et al., 2015; Kurth-Nelson et al., 2016). We have discussed above the ways in which the introduction of mechanisms that replay and learn offline from past experiences can improve the performance of deep RL agents such as DQN (as discussed above in Episodic Memory).

Some encouraging initial progress toward simulation-based planning has been made using deep generative models (Eslami et al., 2016; Rezende et al., 2016a, 2016b) (Figure 2). In particular, recent work has introduced new architectures that have the capacity to generate temporally consistent sequences of generated samples that reflect the geometric layout of newly experienced realistic environments (Gemici et al., 2017; Oh et al., 2015) (Figure 2E), providing a parallel to the function of the hippocampus in binding together multiple components to create an imagined experience that is spatially and temporally coherent (Hassabis and Maguire, 2007). Deep generative models thus show the potential to capture the rich dynamics of complex realistic environments, but using these models for simulation-based planning in agents remains a challenge for future work.

Insights from neuroscience may provide guidance that facilitates the integration of simulation with control. An emerging picture from neuroscience research suggests that the hippocampus supports planning by instantiating an internal model of the environment, with goal-contingent valuation of simulated outcomes occurring in areas downstream of the hippocampus

such the orbitofrontal cortex or striatum (Redish, 2016). Notably, however, the mechanisms that guide the rolling forward of an internal model of the environment in the hippocampus remain uncertain and merit future scrutiny. One possibility is that this process is initiated by the prefrontal cortex through interactions with the hippocampus. Indeed, this notion has distinct parallels with proposals from AI research that a separate controller interacts with an internal model of the environment in a bidirectional fashion, querying the model based on task-relevant goals and receiving predicted simulated states as input (Schmidhuber, 2014). Further, recent efforts to develop agents have employed architectures that instantiate a separation between controller and environmental model to effect simulation-based planning in problems involving the interaction between physical objects (Hamrick et al., 2017).

In enhancing agent capabilities in simulation-based planning, it will also be important to consider other salient properties of this process in humans (Hassabis and Maguire, 2007, 2009). Research into human imagination emphasizes its constructive nature, with humans able to construct fictitious mental scenarios by recombining familiar elements in novel ways, necessitating compositional/disentangled representations of the form present in certain generative models (Eslami et al., 2016; Higgins et al., 2016; Rezende et al., 2016a). This fits well with the notion that planning in humans involves efficient representations that support generalization and transfer, so that plans forged in one setting (e.g., going through a door to reach a room) can be leveraged in novel environments that share structure. Further, planning and mental simulation in humans are "jumpy," bridging multiple temporal scales at a time; for example, humans seem to plan hierarchically, by considering in parallel terminal solutions, interim choice points, and piecemeal steps toward the goal (Balaguer et al., 2016; Solway et al., 2014; Huys et al., 2012). We think that ultimately these flexible, combinatorial aspects of planning will form a critical underpinning of what is perhaps the hardest challenge for AI research: to build an agent that can plan hierarchically, is truly creative, and can generate solutions to challenges that currently elude even the human mind.

Virtual Brain Analytics

One rather different way in which neuroscience may serve AI is by furnishing new analytic tools for understanding computation in AI systems. Due to their complexity, the products of AI research often remain "black boxes"; we understand only poorly the nature of the computations that occur, or representations that are formed, during learning of complex tasks. However, by applying tools from neuroscience to AI systems, synthetic equivalents of single-cell recording, neuroimaging, and lesion techniques, we can gain insights into the key drivers of successful learning in AI research and increase the interpretability of these systems. We call this "virtual brain analytics."

Recent work has made some progress along these lines. For example, visualizing brain states through dimensionality reduction is commonplace in neuroscience, and has recently been applied to neural networks (Zahavy et al., 2016). Receptive field mapping, another standard tool in neuroscience, allows AI researchers to determine the response properties of units in a neural network. One interesting application of this approach in AI is known as activity maximization, in which a network learns to generate synthetic images by maximizing the activity of certain classes of unit (Nguyen et al., 2016; Simonyan et al., 2013). Elsewhere, neuroscience-inspired analyses of linearized networks have uncovered important principles that may be of general benefit in optimizing learning these networks, and understanding the benefits of network depth and representational structure (McClelland and Rogers, 2003; Saxe et al., 2013).

While this initial progress is encouraging, more work is needed. It remains difficult to characterize the functioning of complex architectures such as networks with external memory (Graves et al., 2016). Nevertheless, AI researchers are in the unique position of having ground truth knowledge of all components of the system, together with the potential to causally manipulate individual elements, an enviable scenario from the perspective of experimental neuroscientists. As such, we encourage AI researchers to use approaches from neuroscience to explore properties of network architectures and agents through analysis, visualization, causal manipulation, not forgetting the need for carefully designed hypothesis-driven experiments (Jonas and Kording, 2017; Krakauer et al., 2017). We think that virtual brain analytics is likely to be an increasingly integral part of the pipeline of algorithmic development as the complexity of architectures increases.

From AI to Neuroscience

Thus far, our review has focused primarily on the role of neuroscience in accelerating AI research rather than vice versa. Historically, however, the flow of information between neuroscience and AI has been reciprocal. Machine learning techniques have transformed the analysis of neuroimaging datasets—for example, in the multivariate analysis of fMRI and magnetoencephalographic (MEG) data (Cichy et al., 2014; Çukur et al., 2013; Kriegeskorte and Kievit, 2013)—with promise for expediting connectomic analysis (Glasser et al., 2016), among other techniques. Going further, we believe that building intelligent algorithms has the potential to offer new ideas about the underpinnings of intelligence in the brains of humans and other animals. In particular, psychologists and neuroscientists often have only quite vague notions of the mechanisms that underlie the concepts they study. AI research can help, by formalizing these concepts in a quantitative language and offering insights into their necessity and sufficiency (or otherwise) for intelligent behavior.

A key illustration of this potential is provided by RL. After ideas from animal psychology helped to give birth to reinforcement learning research, key concepts from the latter fed back to inform neuroscience. In particular, the profile of neural signals observed in midbrain dopaminergic neurons in conditioning paradigms was found to bear a striking resemblance to TD-generated prediction errors, providing neural evidence that the brain implements a form of TD learning (O'Doherty et al., 2003; Schultz et al., 1997). This overall narrative arc provides an excellent illustration of how the exchange of ideas between AI and neuroscience can create a "virtuous circle" advancing the objectives of both fields.

In another domain, work focused on enhancing the performance of CNNs has also yielded new insights into the nature of neural representations in high-level visual areas (Khaligh-Razavi and Kriegeskorte, 2014; Yamins and DiCarlo, 2016). For

example, one group systematically compared the ability of more than 30 network architectures from AI to explain the structure of neural representations observed in the ventral visual stream of humans and monkeys, finding favorable evidence for deep supervised networks (Khaligh-Razavi and Kriegeskorte, 2014). Further, these deep convolutional network architectures offer a computational account of recent neurophysiological data demonstrating that the coding of category-orthogonal properties of objects (e.g., position, size) actually increases as one progresses higher up the ventral visual stream (Hong et al., 2016). While these findings are far from definitive as yet, it shows how state-of-the-art neural networks from AI can be used as plausible simulacra of biological brains, potentially providing detailed explanations of the computations occurring therein (Khaligh-Razavi and Kriegeskorte, 2014; Yamins and DiCarlo, 2016). Relatedly, properties of the LSTM architecture have provided key insights that motivated the development of working memory models that afford gating-based maintenance of task-relevant information in the prefrontal cortex (Lloyd et al., 2012; O'Reilly and Frank, 2006).

We also highlight two recent strands of AI research that may motivate new research in neuroscience. First, neural networks with external memory typically allow the controller to iteratively query or "hop through" the contents of memory. This mechanism is critical for reasoning over multiple supporting input statements that relate to a particular query (Sukhbaatar et al., 2015). Previous proposals in neuroscience have argued for a similar mechanism in human cognition, but any potential neural substrates, potentially in the hippocampus, remain to be described (Kumaran and McClelland, 2012). Second, recent work highlights the potential benefits of "meta-reinforcement learning," where RL is used to optimize the weights of a recurrent network such that the latter is able to implement a second, emergent RL algorithm that is able to learn faster than the original (Duan et al., 2016; Wang et al., 2016). Intriguingly, these ideas connect with a growing neuroscience literature indicating a role for the prefrontal cortex in RL, alongside more established dopamine-based mechanisms (Schultz et al., 1997). Specifically, they indicate how a relatively slow-learning dopaminergic RL algorithm may support the emergence of a freestanding RL algorithm instantiated with the recurrent activity dynamics of the prefrontal cortex (Tsutsui et al., 2016).

Insights from AI research are also providing novel perspectives on how the brain might implement an algorithmic parallel to backpropagation, the key mechanism that allows weights within multiple layers of a hierarchical network to be optimized toward an objective function (Hinton et al., 1986; Werbos, 1974). Backpropagation offers a powerful solution to the problem of credit assignment within deep networks, allowing efficient representations to be learned from high dimensional data (LeCun et al., 2015). However, until recently, several aspects of the backpropagation algorithm were viewed to be biologically implausible (e.g., see Bengio et al., 2015). One important factor is that backpropagation has typically been thought to require perfectly symmetric feedback and feedforward connectivity, a profile that is not observed in mammalian brains. Recent work, however, has demonstrated that this constraint can in fact be relaxed (Liao et al., 2015; Lillicrap et al., 2016). Random backward connections, even when held fixed throughout network training, are sufficient to allow the backpropagation algorithm to function effectively through a process whereby adjustment of the forward weights allows backward projections to transmit useful teaching signals (Lillicrap et al., 2016).

A second core objection to the biological plausibility of backpropagation is that weight updates in multi-layered networks require access to information that is non-local (i.e., error signals generated by units many layers downstream) (for review, see Bengio et al., 2015). In contrast, plasticity in biological synapses depends primarily on local information (i.e., pre- and post-synaptic neuronal activity) (Bi and Poo, 1998). AI research has begun to address this fundamental issue. In particular, recent work has shown that hierarchical auto-encoder networks and energy-based networks (e.g., continuous Hopfield networks) (Scellier and Bengio, 2016; Whittington and Bogacz, 2017)—models that have strong connections to theoretical neuroscience ideas about predictive coding (Bastos et al., 2012)—are capable of approximating the backpropagation algorithm, based on weight updates that involve purely local information. Indeed, concrete connections have been drawn between learning in such networks and spike-timing dependent plasticity (Scellier and Bengio, 2016), a Hebbian mechanism instantiated widely across the brain (Bi and Poo, 1998). A different class of local learning rule has been shown to allow hierarchical supervised networks to generate high-level invariances characteristic of biological systems, including mirror-symmetric tuning to physically symmetric stimuli, such as faces (Leibo et al., 2017). Taken together, recent AI research offers the promise of discovering mechanisms by which the brain may implement algorithms with the functionality of backpropagation. Moreover, these developments illustrate the potential for synergistic interactions between AI and neuroscience: research aimed to develop biologically plausible forms of backpropagation have also been motivated by the search for alternative learning algorithms. Given the increasingly deep networks (e.g., >20 layer) used in AI research, factors such as the compounding of successive non-linearities pose challenges for optimization using backpropagation (Bengio et al., 2015).

Conclusions

In this perspective, we have reviewed some of the many ways in which neuroscience has made fundamental contributions to advancing AI research, and argued for its increasingly important relevance. In strategizing for the future exchange between the two fields, it is important to appreciate that the past contributions of neuroscience to AI have rarely involved a simple transfer of full-fledged solutions that could be directly re-implemented in machines. Rather, neuroscience has typically been useful in a subtler way, stimulating algorithmic-level questions about facets of animal learning and intelligence of interest to AI researchers and providing initial leads toward relevant mechanisms. As such, our view is that leveraging insights gained from neuroscience research will expedite progress in AI research, and this will be most effective if AI researchers actively initiate collaborations with neuroscientists to highlight key questions that could be addressed by empirical work.

The successful transfer of insights gained from neuroscience to the development of AI algorithms is critically dependent on the interaction between researchers working in both these fields, with insights often developing through a continual handing back and forth of ideas between fields. In the future, we hope that greater collaboration between researchers in neuroscience and AI, and the identification of a common language between the two fields (Marblestone et al., 2016), will permit a virtuous circle whereby research is accelerated through shared theoretical insights and common empirical advances. We believe that the quest to develop AI will ultimately also lead to a better understanding of our own minds and thought processes. Distilling intelligence into an algorithmic construct and comparing it to the human brain might yield insights into some of the deepest and the most enduring mysteries of the mind, such as the nature of creativity, dreams, and perhaps one day, even consciousness.

ACKNOWLEDGMENTS

We thank Peter Battaglia, Koray Kavukcuoglu, Neil Rabinowitz, Adam Santoro, Greg Wayne, Daan Wierstra, Jane Wang, Martin Chadwick, Joel Leibo, and David Barrett for useful discussions and comments.

REFERENCES

Adolph, K. (2005). Learning to learn in the development of action. In Action as an organizer of perception and cognition during learning and development, J. Lockman, J. Reiser, and C.A. Nelson, eds. (Erlbaum Press), pp. 91–122.

Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., and Qin, Y. (2004). An integrated theory of the mind. Psychol. Rev. *111*, 1036–1060.

Ba, J.L., Mnih, V., and Kavukcuoglu, K. (2015). Multiple object recognition with visual attention. arXiv, arXiv:14127755.

Baddeley, A. (2012). Working memory: theories, models, and controversies. Annu. Rev. Psychol. *63*, 1–29.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv, arXiv:14090473.

Balaguer, J., Spiers, H., Hassabis, D., and Summerfield, C. (2016). Neural Mechanisms of Hierarchical Planning in a Virtual Subway Network. Neuron *90*, 893–903.

Barlow, H. (1959). Sensory mechanisms, the reduction of redundancy, and intelligence. In The Mechanisation of Thought Processes (National Physical Laboratory, UK: H.M. Stationery Office), pp. 535–539.

Barnett, S.M., and Ceci, S.J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. Psychol. Bull. *128*, 612–637.

Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., and Friston, K.J. (2012). Canonical microcircuits for predictive coding. Neuron *76*, 695–711.

Battaglia, P.W., Hamrick, J.B., and Tenenbaum, J.B. (2013). Simulation as an engine of physical scene understanding. Proc. Natl. Acad. Sci. USA *110*, 18327–18332.

Battaglia, P., Pascanu, R., Lai, M., Rezende, D., and Kavukcuoglu, K. (2016). Interaction networks for learning about objects, relations and physics. arXiv, arXiv:161200222.

Bengio, Y., Lee, D.H., Bornschein, J., Mesnard, T., and Lin, Z. (2015). Towards biologically plausible deep learning. arXiv, arXiv:150204156.

Bi, G.Q., and Poo, M.M. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. J. Neurosci. *18*, 10464–10472.

Blundell, C., Uria, B., Pritzel, A., Yazhe, L., Ruderman, A., Leibo, J.Z., Rae, J., Wierstra, D., and Hassabis, D. (2016). Model-free episodic control. arXiv, arXiv:160604460.

Botvinick, M.M., and Plaut, D.C. (2006). Short-term memory for serial order: a recurrent neural network model. Psychol. Rev. *113*, 201–233.

Brooks, R., Hassabis, D., Bray, D., and Shashua, A. (2012). Turing centenary: is the brain a good model for machine intelligence? Nature *482*, 462–463.

Browne, C., Powley, E., Whitehouse, D., Lucas, S.M., Cowling, P.I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S., and Colton, S. (2012). A survey of Monte-Carlo tree search methods. IEEE Trans. Comput. Intell. AI Games *4*, 1–43.

Chang, M.B., Ullman, T., Torralba, A., and Tenenbaum, J.B. (2016). A compositional object-based approach to learning physical dynamics. arXiv, arXiv:161200341.

Chiappa, S., Racaniere, S., Wierstra, D., and Mohamed, S. (2017). Recurrent environment simulators. Proceedings of the 32nd International Conference on Machine Learning, pp. 1–61.

Churchland, P.S., and Sejnowski, T.J. (1988). Perspectives on cognitive neuroscience. Science *242*, 741–745.

Cichon, J., and Gan, W.B. (2015). Branch-specific dendritic Ca(2+) spikes cause persistent synaptic plasticity. Nature *520*, 180–185.

Cichy, R.M., Pantazis, D., and Oliva, A. (2014). Resolving human object recognition in space and time. Nat. Neurosci. *17*, 455–462.

Collins, A., and Koechlin, E. (2012). Reasoning, learning, and creativity: frontal lobe function and human decision-making. PLoS Biol. *10*, e1001293.

Constantinescu, A.O., O'Reilly, J.X., and Behrens, T.E. (2016). Organizing conceptual knowledge in humans with a gridlike code. Science *352*, 1464–1468.

Craik, K. (1943). The Nature of Explanation (Cambridge University Press).

Çukur, T., Nishimoto, S., Huth, A.G., and Gallant, J.L. (2013). Attention during natural vision warps semantic representation across the human brain. Nat. Neurosci. *16*, 763–770.

Daw, N.D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci. *8*, 1704–1711.

Deisseroth, K., and Schnitzer, M.J. (2013). Engineering approaches to illuminating brain structure and dynamics. Neuron *80*, 568–577.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., and Fei-Fei, L. (2009). Imagenet: a large-scale hierarchical image database. In Computer Vision and Pattern Recognition, pp. 1–8.

Denil, M., Agrawal, P., Kulkarni, T.D., Erez, T., Battaglia, P., and de Freitas, N. (2016). Learning to perform physics experiments via deep reinforcement learning. arXiv, arXiv:161101843.

Dolan, R.J., and Dayan, P. (2013). Goals and habits in the brain. Neuron *80*, 312–325.

Doll, B.B., Duncan, K.D., Simon, D.A., Shohamy, D., and Daw, N.D. (2015). Model-based choices involve prospective neural activity. Nat. Neurosci. *18*, 767–772.

Donoso, M., Collins, A.G., and Koechlin, E. (2014). Human cognition. Foundations of human reasoning in the prefrontal cortex. Science *344*, 1481–1486.

Doumas, L.A., Hummel, J.E., and Sandhofer, C.M. (2008). A theory of the discovery and predication of relational concepts. Psychol. Rev. *115*, 1–43.

Duan, Y., Schulman, J., Chen, X., Bartlett, P.L., Sutskever, I., and Abbeel, P. (2016). RL^2: fast reinforcement learning via slow reinforcement learning. arXiv, arXiv:1611.02779.

Durstewitz, D., Seamans, J.K., and Sejnowski, T.J. (2000). Neurocomputational models of working memory. Nat. Neurosci. *3* (*Suppl*), 1184–1191.

Elman, J.L. (1990). Finding structure in time. Cogn. Sci. *14*, 179–211.

Eslami, A., Heess, N., Weber, T.Y.T., Szepesvari, D., Kavukcuoglu, K., and Hinton, G. (2016). Attend, infer, repeat: fast scene understanding with generative models. arXiv, arXiv:160308575.

Esser, S.K., Merolla, P.A., Arthur, J.V., Cassidy, A.S., Appuswamy, R., Andreopoulos, A., Berg, D.J., McKinstry, J.L., Melano, T., Barch, D.R., et al. (2016). Convolutional networks for fast, energy-efficient neuromorphic computing. Proc. Natl. Acad. Sci. USA *113*, 11441–11446.

Fodor, J.A., and Pylyshyn, Z.W. (1988). Connectionism and cognitive architecture: a critical analysis. Cognition *28*, 3–71.

French, R.M. (1999). Catastrophic forgetting in connectionist networks. Trends Cogn. Sci. *3*, 128–135.

Fukushima, K. (1980). Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern. *36*, 193–202.

Fusi, S., Drew, P.J., and Abbott, L.F. (2005). Cascade models of synaptically stored memories. Neuron *45*, 599–611.

Gallistel, C., and King, A.P. (2009). Memory and the Computational Brain: Why Cognitive Science will Transform Neuroscience (Wiley-Blackwell).

Gatys, L.A., Ecker, A.S., and Bethge, M. (2015). A neural algorithm of artistic style. arXiv, arXiv:1508.06576.

Gemici, M., Hung, C., Santoro, A., Wayne, G., Mohamed, S., Rezende, D., Amos, D., and Lillicrap, T. (2017). Generative temporal models with memory. arXiv, arXiv:170204649.

Gershman, S.J., and Daw, N.D. (2017). Reinforcement learning and episodic memory in humans and animals: an integrative framework. Annu. Rev. Psychol. *68*, 101–128.

Gilmore, C.K., McCarthy, S.E., and Spelke, E.S. (2007). Symbolic arithmetic knowledge without instruction. Nature *447*, 589–591.

Glasser, M.F., Smith, S.M., Marcus, D.S., Andersson, J.L., Auerbach, E.J., Behrens, T.E., Coalson, T.S., Harms, M.P., Jenkinson, M., Moeller, S., et al. (2016). The Human Connectome Project's neuroimaging approach. Nat. Neurosci. *19*, 1175–1187.

Goldman-Rakic, P.S. (1990). Cellular and circuit basis of working memory in prefrontal cortex of nonhuman primates. Prog. Brain Res. *85*, 325–335.

Gopnik, A., and Schulz, L. (2004). Mechanisms of theory formation in young children. Trends Cogn. Sci. *8*, 371–377.

Graves, A., Wayne, G., and Danihelka, I. (2014). Neural turing machines. arXiv, arXiv:1410.5401.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S.G., Grefenstette, E., Ramalho, T., Agapiou, J., et al. (2016). Hybrid computing using a neural network with dynamic external memory. Nature *538*, 471–476.

Gregor, K., Danihelka, I., Graves, A., Renzende, D., and Wierstra, D. (2015). DRAW: a recurrent neural network for image generation. arXiv, arXiv:150204623.

Hafner, R., and Riedmiller, M. (2011). Reinforcement learning in feedback control. Mach. Learn. *84*, 137–169.

Hamrick, J.B., Ballard, A.J., Pascanu, R., Vinyals, O., Heess, N., and Battaglia, P.W. (2017). Metacontrol for adaptive imagination-based optimization. Proceedings of the 5th International Conference on Learning Representations (ICLR 2017), pp. 1–21.

Harlow, H.F. (1949). The formation of learning sets. Psychol. Rev. *56*, 51–65.

Hassabis, D., and Maguire, E.A. (2007). Deconstructing episodic memory with construction. Trends Cogn. Sci. *11*, 299–306.

Hassabis, D., and Maguire, E.A. (2009). The construction system of the brain. Philos. Trans. R. Soc. Lond. B Biol. Sci. *364*, 1263–1271.

Haugeland, J. (1985). Artificial Intelligence: The Very Idea (MIT Press).

Hayashi-Takagi, A., Yagishita, S., Nakamura, M., Shirai, F., Wu, Y.I., Loshbaugh, A.L., Kuhlman, B., Hahn, K.M., and Kasai, H. (2015). Labelling and optical erasure of synaptic memory traces in the motor cortex. Nature *525*, 333–338.

Hebb, D.O. (1949). The Organization of Behavior (John Wiley & Sons).

Higgins, I., Matthey, L., Glorot, X., Pal, A., Uria, B., Blundell, C., Mohamed, S., and Lerchner, A. (2016). Early visual concept learning with unsupervised deep learning. arXiv, arXiv:160605579.

Hinton, G.E., McClelland, J.L., and Rumelhart, D.E. (1986). Distributed Representations. In Explorations in the Microstructure of Cognition (MIT Press), pp. 77–109.

Hinton, G.E., Osindero, S., and Teh, Y.W. (2006). A fast learning algorithm for deep belief nets. Neural Comput. *18*, 1527–1554.

Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. arXiv, arXiv:12070580.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. Neural Comput. *9*, 1735–1780.

Holyoak, K.J., and Thagard, P. (1997). The analogical mind. Am. Psychol. *52*, 35–44.

Hong, S., Oh, J., Bohyung, H., and Lee, H. (2015). Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. arXiv, arXiv:151207928.

Hong, H., Yamins, D.L., Majaj, N.J., and DiCarlo, J.J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. Nat. Neurosci. *19*, 613–622.

Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. USA *79*, 2554–2558.

Hopfield, J.J., and Tank, D.W. (1986). Computing with neural circuits: a model. Science *233*, 625–633.

Hubel, D.H., and Wiesel, T.N. (1959). Receptive fields of single neurones in the cat's striate cortex. J. Physiol. *148*, 574–591.

Huys, Q.J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., and Roiser, J.P. (2012). Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. PLoS Comput. Biol. *8*, e1002410.

Johnson, A., and Redish, A.D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. J. Neurosci. *27*, 12176–12189.

Jonas, E., and Kording, K.P. (2017). Could a neuroscientist understand a microprocessor? PLoS Comput. Biol. *13*, e1005268.

Jordan, M.I. (1997). Serial order: a parallel distributed processing approach. Adv. Psychol. *121*, 471–495.

Kemp, C., Goodman, N.D., and Tenenbaum, J.B. (2010). Learning to learn causal models. Cogn. Sci. *34*, 1185–1243.

Khaligh-Razavi, S.M., and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput. Biol. *10*, e1003915.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. Proc. Natl. Acad. Sci. USA *114*, 3521–3526.

Koch, C., and Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. Hum. Neurobiol. *4*, 219–227.

Krakauer, J.W., Ghazanfar, A.A., Gomez-Marin, A., MacIver, M.A., and Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. Neuron *93*, 480–490.

Kriegeskorte, N., and Kievit, R.A. (2013). Representational geometry: integrating cognition, computation, and the brain. Trends Cogn. Sci. *17*, 401–412.

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems 25, pp. 1097–1105.

Kumaran, D., and McClelland, J.L. (2012). Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system. Psychol. Rev. *119*, 573–616.

Kumaran, D., Hassabis, D., and McClelland, J.L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. Trends Cogn. Sci. *20*, 512–534.

Kurth-Nelson, Z., Economides, M., Dolan, R.J., and Dayan, P. (2016). Fast sequences of non-spatial state representations in humans. Neuron *91*, 194–204.

Lake, B.M., Salakhutdinov, R., and Tenenbaum, J.B. (2015). Human-level concept learning through probabilistic program induction. Science *350*, 1332–1338.

Lake, B.M., Ullman, T.D., Tenenbaum, J.B., and Gershman, S.J. (2016). Building machines that learn and think like people. arXiv, arXiv:1604.00289.

Larochelle, H., and Hinton, G. (2010). Learning to combine foveal glimpses with a third-order Boltzmann machine. NIPS'10 Proceedings of the International Conference on Neural Information Processing Systems, pp. 1243–1251.

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., and Jackel, L.D. (1989). Backpropagation applied to handwritten zip code recognition. Neural Comput. *1*, 541–551.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444.

Leibo, J.Z., Liao, Q., Anselmi, F., Freiwald, W.A., and Poggio, T. (2017). View-tolerant face recognition and Hebbian Learning imply mirror-symmetric neural tuning to head orientation. Curr. Biol. *27*, 62–67.

Legg, S., and Hutter, M. (2007). A collection of definitions of intelligence. In Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms—Proceedings of the AGI Workshop, B. Goertzel and P. Wang, eds. (Amsterdam IOS), pp. 17–24.

Lengyel, M., and Dayan, P. (2007). Hippocampal contributions to control: the third way. In Advances in Neural Information Processing Systems 20, pp. 889–896.

Liao, Q., Leibo, J.Z., and Poggio, T. (2015). How important is weight symmetry in backpropagation? arXiv, arXiv:151005067.

Lillicrap, T.P., Cownden, D., Tweed, D.B., and Akerman, C.J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. Nat. Commun. *7*, 13276.

Lloyd, K., Becker, N., Jones, M.W., and Bogacz, R. (2012). Learning to use working memory: a reinforcement learning gating model of rule acquisition in rats. Front. Comput. Neurosci. *6*, 87.

Marblestone, A.H., Wayne, G., and Kording, K.P. (2016). Toward an integration of deep learning and neuroscience. Front. Comput. Neurosci. *10*, 94.

Marcus, G.F. (1998). Rethinking eliminative connectionism. Cognit. Psychol. *37*, 243–282.

Markram, H. (2006). The blue brain project. Nat. Rev. Neurosci. *7*, 153–160.

Marr, D., and Poggio, T. (1976). From understanding computation to understanding neural circuitry. A.I. Memo *357*, 1–22.

McClelland, J.L., and Rogers, T.T. (2003). The parallel distributed processing approach to semantic cognition. Nat. Rev. Neurosci. *4*, 310–322.

McClelland, J.L., McNaughton, B.L., and O'Reilly, R.C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychol. Rev. *102*, 419–457.

McCulloch, W., and Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. Bull. Math. Biophys. *5*, 115–133.

Mnih, V., Heess, N., Graves, A., and Kavukcuoglu, K. (2014). Recurrent models of visual attention. arXiv, arXiv:14066247.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. Nature *518*, 529–533.

Moore, T., and Zirnsak, M. (2017). Neural mechanisms of selective visual attention. Annu. Rev. Psychol. *68*, 47–72.

Moravčík, M., Schmid, M., Burch, N., Lisý, V., Morrill, D., Bard, N., Davis, T., Waugh, K., Johanson, M., and Bowling, M. (2017). DeepStack: expert-level artificial intelligence in heads-up no-limit poker. Science *356*, 508–513.

Nguyen, A., Dosovitskiy, A., Yosinski, J., Borx, T., and Clune, J. (2016). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. arXiv, arXiv:160509304.

Nishiyama, J., and Yasuda, R. (2015). Biochemical computation for spine structural plasticity. Neuron *87*, 63–75.

O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H., and Dolan, R.J. (2003). Temporal difference models and reward-related learning in the human brain. Neuron *38*, 329–337.

O'Neill, J., Pleydell-Bouverie, B., Dupret, D., and Csicsvari, J. (2010). Play it again: reactivation of waking experience and memory. Trends Neurosci. *33*, 220–229.

O'Reilly, R.C., and Frank, M.J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. Neural Comput. *18*, 283–328.

Oh, J., Guo, X., Lee, H., Lewis, R., and Singh, S. (2015). Action-conditional video prediction using deep networks in Atari games. arXiv, arXiv:150708750.

Ólafsdóttir, H.F., Barry, C., Saleem, A.B., Hassabis, D., and Spiers, H.J. (2015). Hippocampal place cells construct reward related sequences through unexplored space. eLife *4*, e06063.

Olshausen, B.A., Anderson, C.H., and Van Essen, D.C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. J. Neurosci. *13*, 4700–4719.

Pfeiffer, B.E., and Foster, D.J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. Nature *497*, 74–79.

Posner, M.I., and Petersen, S.E. (1990). The attention system of the human brain. Annu. Rev. Neurosci. *13*, 25–42.

Raby, C.R., Alexis, D.M., Dickinson, A., and Clayton, N.S. (2007). Planning for the future by western scrub-jays. Nature *445*, 919–921.

Redish, A.D. (2016). Vicarious trial and error. Nat. Rev. Neurosci. *17*, 147–159.

Reed, S., Zhang, Y., Zhang, Y., and Lee, S. (2015). Deep visual analogy-making. In NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems, pp. 1252–1260.

Reed, S., Akata, Z., Mohan, S., Tenka, S., Schiele, B., and Lee, H. (2016). Learning what and where to draw. arXiv, arXiv:161002454.

Rezende, D., Eslami, A., Mohamed, S., Battaglia, P., Jaderberg, M., and Heess, N. (2016a). Unsupervised learning of 3D structure from images. arXiv, arXiv:160700662.

Rezende, D., Mohamed, S., Danihelka, I., Gregor, K., and Wierstra, D. (2016b). One-shot generalization in deep generative models. arXiv, arXiv:160305106.

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nat. Neurosci. *2*, 1019–1025.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. Psychol. Rev. *65*, 386–408.

Rowland, D.C., Roudi, Y., Moser, M.B., and Moser, E.I. (2016). Ten years of grid cells. Annu. Rev. Neurosci. *39*, 19–40.

Rumelhart, D.E., Hinton, G., and Williams, R.J. (1985). Learning internal representations by error propagation. In Parallel Distributed Processing: Explorations in the Microstructure of Cognition, *Volume 1* (MIT Press), pp. 318–362.

Rumelhart, D.E., McClelland, J.L., and Group, P.R. (1986). Parallel Distributed Processing: Explorations in the Microstructures of Cognition, *Volume 1* (MIT Press).

Rusu, A.A., Rabinowitz, N., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., and Hadsell, R. (2016a). Progressive neural networks. arXiv, arXiv:160604671.

Rusu, A.A., Vecerik, M., Rothorl, T., Heess, N., Pascanu, R., and Hadsell, R. (2016b). Sim-to-real robot learning from pixels with progressive nets. arXiv, arXiv:161004286.

Salinas, E., and Abbott, L.F. (1997). Invariant visual responses from attentional gain fields. J. Neurophysiol. *77*, 3267–3272.

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., and Lillicrap, T. (2016). One-shot Learning with Memory-Augmented Neural Networks. arXiv, arXiv:160506065.

Santoro, A., Raposo, D., Barrett, D.G.T., Malinowski, M., Pascanu, R., Battaglia, P., and Lillicrap. T. (2017). A simple neural network module for relational reasoning. arXiv, arXiv:1706.01427, https://arxiv.org/abs/1706.01427.

Saxe, A.M., Ganguli, S., and McClelland, J.L. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv, arXiv:13126120v3.

Scellier, B., and Bengio, Y. (2016). Equilibrium propagation: bridging the gap between energy-based models and backpropagation. arXiv, arXiv:160205179.

Schacter, D.L., Addis, D.R., Hassabis, D., Martin, V.C., Spreng, R.N., and Szpunar, K.K. (2012). The future of memory: remembering, imagining, and the brain. Neuron *76*, 677–694.

Schaul, T., Quan, J., Antonoglou, I., and Silver, D. (2015). Prioritized experience replay. bioRxiv, arXiv:1511.05952.

Schmidhuber, J. (2014). Deep learning in neural networks: an overview. arXiv, 14047828.

Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. Science *275*, 1593–1599.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. IEEE Trans. Pattern Anal. Mach. Intell. *29*, 411–426.

Shallice, T. (1988). From Neuropsychology to Mental Structure (Cambridge University Press).

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of Go with deep neural networks and tree search. Nature *529*, 484–489.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv, arXiv:13126034.

Singer, A.C., and Frank, L.M. (2009). Rewarded outcomes enhance reactivation of experience in the hippocampus. Neuron *64*, 910–921.

Skaggs, W.E., and McNaughton, B.L. (1996). Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. Science *271*, 1870–1873.

Smith, L.B. (1995). Self-organizing processes in learning to learn words: development is not induction. In Basic and Applied Perspectives on Learning, Cognition, and Development. The Minnesota Symposia on Child Psychology (Lawrence Erlbaum Associates), pp. 1–32.

Solway, A., Diuk, C., Córdova, N., Yee, D., Barto, A.G., Niv, Y., and Botvinick, M.M. (2014). Optimal behavioral hierarchy. PLoS Comput. Biol. *10*, e1003779.

Spelke, E.S., and Kinzler, K.D. (2007). Core knowledge. Dev. Sci. *10*, 89–96.

Squire, L.R., Stark, C.E., and Clark, R.E. (2004). The medial temporal lobe. Annu. Rev. Neurosci. *27*, 279–306.

St. John, M.F., and McClelland, J.L. (1990). Learning and applying contextual constraints in sentence comprehension. Artif. Intell. *46*, 217–257.

Stachenfeld, K., Botvinick, M.M., and Gershman, S.J. (2014). Design principles of hippocampal cognitive maps. In Advances in Neural Information Processing Systems 27, pp. 2528–2536.

Sukhbaatar, S., Szlam, A., Weston, J., and Fergus, R. (2015). End-to-end memory networks. arXiv, arXiv:150308895.

Summerfield, J.J., Lepsien, J., Gitelman, D.R., Mesulam, M.M., and Nobre, A.C. (2006). Orienting attention based on long-term memory experience. Neuron *49*, 905–916.

Sutton, R.S. (1991). Dyna, an integrated architecture for learning, planning, and reacting. ACM SIGART Bull. *2*, 160–163.

Sutton, R.S., and Barto, A.G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. Psychol. Rev. *88*, 135–170.

Sutton, R., and Barto, A. (1998). Reinforcement Learning (MIT Press).

Tesauro, G. (1995). Temporal difference learning and TD-Gammon. Commun. ACM *38*, 58–68.

Thrun, S., and Mitchell, T.M. (1995). Lifelong robot learning. Robot. Auton. Syst. *15*, 25–46.

Tolman, E.C. (1948). Cognitive maps in rats and men. Psychol. Rev. *55*, 189–208.

Tsutsui, K., Grabenhorst, F., Kobayashi, S., and Schultz, W. (2016). A dynamic code for economic object valuation in prefrontal cortex neurons. Nat. Commun. *7*, 12554.

Tulving, E. (1985). How many memory systems are there? American Psychologist *40*, 385–398.

Tulving, E. (2002). Episodic memory: from mind to brain. Annu. Rev. Psychol. *53*, 1–25.

Turing, A.M. (1936). On computable numbers, with an application to the Entscheidungs problem. Proc. Lond. Math. Soc. *2*, 230–265.

Turing, A. (1950). Computing machinery and intelligence. Mind *236*, 433–460.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). WaveNet: a generative model for raw audio. arXiv, arXiv:160903499.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., and Wierstra, D. (2016). Matching networks for one shot learning. arXiv, arXiv:160604080.

Wang, J., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J.Z., Munos, R., Blundell, C., Kumaran, D., and Botvinick, M.M. (2016). Learning to reinforcement learn. arXiv, arXiv:161105763.

Werbos, P.J. (1974). Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences (Harvard University).

Weston, J., Chopra, S., and Bordes, A. (2014). Memory networks. arXiv, arXiv:14103916.

Whittington, J.C.R., and Bogacz, R. (2017). An approximation of the error backpropagation algorithm in a predictive coding network with local Hebbian synaptic plasticity. Neural Comput. *29*, 1229–1262.

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: bridging the gap between human and machine translation. arXiv, arXiv:160908144.

Xu, K., Kiros, J., Courville, A., Salakhutdinov, R., and Bengio, Y. (2015). Show, attend and tell: neural image caption generation with visual attention. arXiv, arXiv:150203044.

Yamins, D.L., and DiCarlo, J.J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nat. Neurosci. *19*, 356–365.

Yang, G., Pan, F., and Gan, W.B. (2009). Stably maintained dendritic spines are associated with lifelong memories. Nature *462*, 920–924.

Zahavy, T., Zrihem, N.B., and Mannor, S. (2016). Graying the black box: understanding DQNs. arXiv, arXiv:160202658.

Zaremba, W., and Sutskever, I. (2014). Learning to execute. arXiv, arXiv:1410.4615.

MACHINE LEARNING AND LAW

Harry Surden*

INTRODUCTION

What impact might artificial intelligence (AI) have upon the practice of law? According to one view, AI should have little bearing upon legal practice barring significant technical advances.[1] The reason is that legal practice is thought to require advanced cognitive abilities, but such higher-order cognition remains outside the capability of current AI technology.[2] Attorneys, for example, routinely combine abstract reasoning and problem solving skills in environments of legal and factual uncertainty.[3] Modern AI algorithms, by contrast, have been unable to replicate most human intellectual abilities, falling far short in advanced cognitive processes—such as analogical reasoning—that are basic to legal practice.[4] Given these and other limitations in current AI technology, one might conclude that until computers can replicate the higher-order cognition routinely displayed by trained attorneys, AI would have little impact in a domain as full of abstraction and uncertainty as law.[5]

Although there is some truth to that view, its conclusion is overly broad. It misses a class of legal tasks for which current AI technology

---

* Associate Professor of Law, University of Colorado Law School; B.A. Cornell University; J.D. Stanford University; Affiliated Faculty, Stanford Codex Center for Legal Informatics. I would like to thank my colleagues at the University of Colorado for their insightful comments, and Ted Sichelman, Seema Shah, and Dan Katz for their helpful observations and suggestions.

1. *See, e.g.*, Symposium, *Legal Reasoning and Artificial Intelligence: How Computers "Think" Like Lawyers*, 8 U. CHI. L. SCH. ROUNDTABLE 1, 19 (2001) (Cass Sunstein argues that, "[A]t the present state of the art artificial intelligence cannot engage in analogical reasoning or legal reasoning").

2. *See, e.g.*, Karl Okamoto, *Teaching Transactional Lawyering*, 1 DREXEL L. REV. 69, 83 (2009) ("The essence of lawyering is 'creative problem solving' under conditions of uncertainty and complexity. This conception of lawyering as problem solving has become commonplace.").

3. *Id.* at 83.

4. *Id.*

5. *See* Harry Surden, *Computable Contracts*, 46 U.C. DAVIS L. REV. 629, 646 (2012) (discussing how language changes that are typically trivial for humans to decipher may confuse computer algorithms).

can still have an impact even given the technological inability to match human-level reasoning. Consider that outside of law, non-cognitive AI techniques have been successfully applied to tasks that were once thought to necessitate human intelligence—for example language translation.[6] While the results of these automated efforts are sometimes imperfect, the interesting point is that such computer generated results have often proven useful for particular tasks where strong approximations are acceptable.[7] In a similar vein, this Article will suggest that there may be a limited, but not insignificant, subset of legal tasks that are capable of being partially automated using current AI techniques despite their limitations relative to human cognition.

In particular, this Article focuses upon a class of AI methods known as "machine learning" techniques and their potential impact upon legal practice. Broadly speaking, machine learning involves computer algorithms that have the ability to "learn" or improve in performance over time on some task.[8] Given that there are multiple AI approaches, why highlight machine learning in particular? In the last few decades, researchers have successfully used machine learning to automate a variety of sophisticated tasks that were previously presumed to require human cognition. These applications range from autonomous (i.e., self-driving) cars, to automated language translation, prediction, speech recognition, and computer vision.[9] Researchers have also begun to apply these techniques in the context of law.[10]

To be clear, I am not suggesting that all, or even most, of the tasks routinely performed by attorneys are automatable given the current state of AI technology. To the contrary, many of the tasks performed by attorneys do appear to require the type of higher order intellectual skills that are beyond the capability of current techniques. Rather, I am suggesting that there are subsets of legal tasks that are likely

---

6. *See* DAVID BELLOS, IS THAT A FISH IN YOUR EAR?: TRANSLATION AND THE MEANING OF EVERYTHING 253–57 (2011); *Find Out How Our Translations Are Created*, GOOGLE, http://translate.google.com/about (last visited Feb. 24, 2014).

7. *See* BELLOS, *supra* note 6.

8. PETER FLACH, MACHINE LEARNING: THE ART AND SCIENCE OF ALGORITHMS THAT MAKE SENSE OF DATA 3 (2012).

9. Burkhard Bilger, *Auto Correct: Has the Self-Driving Car at Last Arrived?*, NEW YORKER, Nov. 25, 2013, at 96, 106; PARAG KULKARNI, REINFORCEMENT AND SYSTEMIC MACHINE LEARNING FOR DECISION MAKING 1–2 (2012) (discussing computer vision).

10. *See, e.g.*, Daniel Martin Katz, *Quantitative Legal Prediction—or—How I Learned to Stop Worrying and Start Preparing for the Data-Driven Future of the Legal Services Industry*, 62 EMORY L.J. 909, 936 (2013) (discussing legal applications such as automation in document discovery and quantitative legal prediction).

automatable under the current state of the art, provided that the technologies are appropriately matched to relevant tasks, and that accuracy limitations are understood and accounted for. In other words, even given current limitations in AI technology as compared to human cognition, such computational approaches to automation may produce results that are "good enough" in certain legal contexts.

Part I of this Article explains the basic concepts underlying machine learning. Part II will convey a more general principle: non-intelligent computer algorithms can sometimes produce intelligent results in complex tasks through the use of suitable proxies detected in data. Part III will explore how certain legal tasks might be amenable to partial automation under this principle by employing machine learning techniques. This Part will also emphasize the significant limitations of these automated methods as compared to the capabilities of similarly situated attorneys.

I. OVERVIEW OF MACHINE LEARNING

A. *What Is Machine Learning?*

"Machine learning" refers to a subfield of computer science concerned with computer programs that are able to learn from experience and thus improve their performance over time.[11] As will be discussed, the idea that the computers are "learning" is largely a metaphor and does not imply that computers systems are artificially replicating the advanced cognitive systems thought to be involved in human learning.[12] Rather, we can consider these algorithms to be learning in a *functional* sense: they are capable of changing their behavior to enhance their performance on some task through experience.[13]

Commonly, machine learning algorithms are used to detect patterns in data in order to automate complex tasks or make predictions.[14] Today, such algorithms are used in a variety of real-world commercial applications including Internet search results, facial recognition, fraud

---

11. STUART RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 693 (3d ed. 2010).

12. I. H. WITTEN, DATA MINING: PRACTICAL MACHINE LEARNING TOOLS AND TECHNIQUES § 1.3 (3d ed. 2011).

13. *Id.*

14. David E. Sorkin, *Technical and Legal Approaches to Unsolicited Electronic Mail*, 35 U.S.F. L. REV. 325, 326 (2001).

detection, and data mining.[15] Machine learning is closely associated with the larger enterprise of "predictive analytics" as researchers often employ machine learning methods to analyze existing data to predict the likelihood of uncertain outcomes.[16] If performing well, machine learning algorithms may produce automated results that approximate those that would have been made by a similarly situated person. Machine learning is thus often considered a branch of artificial intelligence, since a well-performing algorithm may produce automated results that appear "intelligent."[17]

The goal of this Part is to convey some basic principles of machine learning in a manner accessible to non-technical audiences in order to express a larger point about the potential applicability of these techniques to tasks within the law.

1. Email Spam Filters as an Example of Machine Learning

Consider a familiar example—email "spam" filters—that will illustrate some basic features common to machine learning techniques. "Spam" emails are unsolicited, unwanted commercial emails that can interfere with a user accessing more important communications.[18] In principle, an email user could manage spam manually by reading each email, identifying whether a given email is spam, and deleting those determined to be spam. However, given that this task is labor intensive, it would be desirable to automate spam identification. To perform such automated filtering of spam, email software programs frequently use machine learning algorithms.[19]

How do machine learning algorithms automatically identify spam? Such algorithms are designed to detect patterns among data. In a typical process, a machine learning algorithm is "trained" to recognize spam emails by providing the algorithm with known examples of spam for pattern analysis. For instance, imagine that a person determines that a particular email is spam and flags it as such using her email reading software. We can think of this act of flagging as an indication to the computer algorithm that this is a verified example of a spam email that

---

15. WITTEN, *supra* note 12, at § 1.3.

16. *See, e.g.*, LAWRENCE MAISEL, PREDICTIVE BUSINESS ANALYTICS: FORWARD LOOKING CAPABILITIES TO IMPROVE BUSINESS PERFORMANCE, 27–30 (2014).

17. RUSSEL & NORVIG, *supra* note 11, at 3–5.

18. Sorkin, *supra* note 14, at 325–30.

19. *Id.*

should be assessed for patterns.[20]

In analyzing the spam email, the machine learning algorithm will attempt to detect the telltale characteristics that indicate that a given email is more likely than not to be spam. After analyzing several such examples, the algorithm may detect a pattern and infer a general "rule"[21]—for instance that emails with the phrase "Earn Extra Cash" tend to be statistically more likely to be spam emails than wanted emails. It can then use such learned indicia to make automated assessments about the likelihood that a new incoming email is or is not spam.[22]

In general, machine learning algorithms are able to automatically build such heuristics by inferring information through pattern detection in data. If these heuristics are correct, they will allow the algorithm to make predictions or automated decisions involving future data.[23] Here, the algorithm has detected a pattern within the data provided (i.e., the set of example spam emails) that, of the emails that were flagged as spam, many of them contained the phrase "Earn Extra Cash." From this pattern, it then inferred a heuristic: that emails with the text "Earn Extra Cash" were more likely to be spam. Such a generalization can thus be applied going forward to automatically categorize new incoming emails containing "Earn Extra Cash" as spam. The algorithm will attempt to detect other similar patterns that are common among spam emails that can be used as a heuristic for distinguishing spam from wanted emails.

Importantly, machine learning algorithms are designed to improve in performance over time on a particular task as they receive more data. The goal of such an algorithm is to build an internal computer model of some complex phenomenon—here spam emails—that will ultimately allow the computer to make automated, accurate classification decisions.

---

20. In many cases, machine learning algorithms are trained through carefully validated training sets of data, in which the data has been carefully screened and categorized by people. *See, e.g.*, DAVID BARBER, BAYESIAN REASONING AND MACHINE LEARNING 290–96 (2011).

21. The term "rule" is used approximately in the sense of "rule of thumb." This is important, because machine learning is an *inductive* rather a *deductive* technique. In a deductive approach, general logical rules (statements) characterizing the state of the world are expressly articulated, and information is extracted by combining statements according to logical operations. By contrast, in an *inductive* approach, models of the world are developed upon observing the past and expressing the state of the world (often) in probabilities induced from observation, rather than as general rules. *See generally* Katz, *supra* note 10, at 946.

22. To be clear, this is an extreme over-simplification of machine learning for illustrative purposes. Moreover, there are many different machine learning algorithmic strategies other than the particular one illustrated here. *See generally* MEHRYAR MOHRI ET AL., FOUNDATIONS OF MACHINE LEARNING (2012).

23. TOBY SEGARAN, PROGRAMMING COLLECTIVE INTELLIGENCE: BUILDING SMART WEB 2.0 APPLICATIONS 3 (2007).

In this case, the internal model would include multiple rules of thumb about the likely characteristics of spam induced over time—in addition to the "Earn Extra Cash" heuristic just described—that the computer can subsequently follow to classify new, incoming emails.

For instance, such an algorithm might infer from additional spam examples that emails that originate from the country Belarus[24] tend to be more likely to be spam than emails from other countries. Similarly, the algorithm might learn that emails sent from parties that the reader has previously corresponded with are less likely to be spam than those from complete strangers. These additional heuristics that the algorithm learned from analyzing additional data will allow it to make better automated decisions about what is or is not spam.

As illustrated, the rule sets[25] that form the internal model are inferred by examining and detecting patterns within data. Because of this, such rule-sets tend to be built cumulatively over time as more data arrives. Machine learning algorithms typically develop heuristics incrementally by examining each new example and comparing it against prior examples to identify overall commonalities that can be generalized more broadly. For example, an algorithm may have to analyze several thousand examples of spam emails before it detects a reliable pattern such that the text "Earn Extra Cash" is a statistical indicia of likely spam.

For this reason, a machine learning algorithm may perform poorly at first when it has only had a few examples of a phenomenon (e.g., spam emails) from which to detect relevant patterns. At such an early point, its internal rule-set will likely be fairly underdeveloped. However, the ability to detect useful patterns tends to improve as the algorithm is able to examine more examples of the phenomenon at issue. Often, such an algorithm will need data with many hundreds or thousands examples of the relevant phenomenon in order to produce a useful internal model (i.e. robust set of predictive computer rules).[26]

The prior example illustrates what is meant by "learning" in the machine learning context: it is this ability to improve in performance by detecting new or better patterns from additional data. A machine

---

24. *See* Paul Ducklin, *Dirty Dozen Spam Sending Nations*, NAKED SECURITY (Oct. 17, 2013), http://nakedsecurity.sophos.com/2013/10/17/dirty-dozen-spam-sending-nations-find-where-you-finished-in-our-q3-spampionship-chart/.

25. It is important to note that these rule-sets are often actually mathematical functions or some other data structure representing the object to be modeled, rather than a series of formal, general rules. *See* KULKARNI, *supra* note 9, at 2–10.

26. CHRISTOPHER D. MANNING, INTRODUCTION TO INFORMATION RETRIEVAL 335 (2008).

learning algorithm can become more accurate at a task (like classifying email as spam) over time because its design enables it to continually refine its internal model by analyzing more examples and inferring new, useful patterns from additional data.

This capability to improve in performance over time by continually analyzing data to detect additional useful patterns is the key attribute that characterizes machine learning algorithms. Upon the basis of such an incrementally produced model, a well-performing machine learning algorithm may be able to automatically perform a task—such as classifying incoming emails as either spam or wanted emails—with a high degree of accuracy that approximates the classifications that a similarly situated human reviewer would have made.[27]

2. *Detecting Patterns to Model Complex Phenomena*

There are a few points to emphasize about the above example. First, machine learning often (but not exclusively) involves learning from a set of verified examples of some phenomenon. Thus, in the prior example, the algorithm was explicitly provided with a series of emails that a human predetermined to be spam, and learned the characteristics of spam by analyzing these provided examples. This approach is known as "supervised" learning, and the provided examples upon which the algorithm is being trained to recognize patterns are known as the "training set."[28] The goal of such training is to allow the algorithm to create an internal computer model of a given phenomenon that can be generalized to apply to new, never-before-seen examples of that phenomenon.

Second, such machine learning algorithms are able to automatically build accurate models of some phenomenon—here the characteristics of spam email—without being explicitly programmed.[29] Most software is developed by a manual approach in which programmers explicitly specify a series of rules for a computer to follow that will produce some desired behavior. For instance, if designing a spam filter by this manual method, a programmer might first consider the features that she believed to be characteristic of spam, and then proceed to program a computer

---

27. WILLIAM S. YERAZUNIS, THE SPAM-FILTERING ACCURACY PLATEAU AT 99.9 PERCENT ACCURACY AND HOW TO GET PAST IT (Dec. 2004), *available at* http://www.merl.com/reports/docs/TR2004-091.pdf (noting that many spam filters have achieved accuracy rates at over 99.9%).

28. FLACH, *supra* note 8, at 2.

29. Pedro Domingos, *A Few Useful Things to Know About Machine Learning*, COMM. ACM, Oct. 2012, at 80.

with a series of corresponding rules to make automated distinctions.

However, many phenomena are so complicated and dynamic that it is difficult to model them manually.[30] The problem with a manual, bottom-up approach to modeling complex and changing phenomenon (such as spam) is that it is very difficult to specify a rule set ex-ante that would be robust and accurate enough to direct a computer to make useful, automated decisions. For instance, a programmer might not think to include a rule that an email with a Belarus origin should be considered somewhat more likely to be spam. It is often difficult to explicitly program a set of computer rules to produce useful automation when dealing with complex, changing phenomenon.

Machine learning algorithms, by contrast, are able to incrementally build complex models by automatically detecting patterns as data arrives. Such algorithms are powerful because, in a sense, these algorithms program themselves over time with the rules to accomplish a task, rather than being programmed manually with a series of pre-determined rules.[31] The rules are inferred from analyzed data and the model builds itself as additional data is analyzed. For instance, in the above example, as the algorithm encountered new examples of spam with different features, it was able to add to its internal model additional markers of spam that it was able to detect (e.g., emails originating from Belarus). Such an incremental, adaptive, and iterative process often allows for the creation of nuanced models of complex phenomena that may otherwise be too difficult for programmers to specify manually, up front.[32]

Third, what made the discussed spam filtering algorithm a machine *learning* algorithm was that it was able to *improve* its accuracy in classifying spam as it received more examples to analyze. In this sense, we are using a functional meaning of "learning." The algorithms are not learning in the cognitive sense typically associated with human learning. Rather, we can think of the algorithms as learning in the sense that they are changing their behavior to perform better in the future as they receive more data.[33] Thus, in the above example, if the spam filter

---

30. *Id.*

31. TOM MITCHELL, THE DISCIPLINE OF MACHINE LEARNING, REPORT NO. ML-06-CMU-108 § 1 (2006), *available at* http://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf ("Machine Learning focuses on . . . how to get computers to program themselves (from experience plus some initial structure).").

32. *Id.* ("[S]peech recognition accuracy is greater if one trains the system, than if one attempts to program it by hand.").

33. I. H. WITTEN, DATA MINING: PRACTICAL MACHINE LEARNING TOOLS AND TECHNIQUES 8 (2d ed. 2005).

algorithm became more accurate at identifying spam as it received more examples of spam and refined its internal rule-set. We can conceptualize this shift as "learning" from a functional perspective in an analogous way that we often associate human learning with improved performance on some task.

Fourth, the filtering algorithm described used statistical techniques to classify spam. Machine learning algorithms are often (although not exclusively) statistical in nature. Thus, in one sense, machine learning is not very different from the numerous statistical techniques already widely used within empirical studies in law.[34] One salient distinction is that while many existing statistical approaches involve fixed or slow-to-change statistical models, the focus in machine learning is upon computer algorithms that are expressly designed to be dynamic and capable of changing and adapting to new and different circumstances as the data environment shifts.

II. INTELLIGENT RESULTS WITHOUT INTELLIGENCE

A. Proxies and Heuristics for Intelligence

The prior example was meant to illustrate a broader point: one can sometimes accomplish tasks associated with human intelligence with non-intelligent computer algorithms. There are certain tasks that appear to require intelligence because when humans perform them, they implicate higher-order cognitive skills such as reasoning, comprehension, meta-cognition, or contextual perception of abstract concepts. However, research has shown that certain of these tasks can be automated—to some degree—through the use of non-cognitive computational techniques that employ heuristics or proxies (e.g., statistical correlations) to produce useful, "intelligent" results. By a proxy or heuristic, I refer to something that is an effective stand-in for some underlying concept, feature, or phenomenon.

To say it differently, non-cognitive computer algorithms can sometimes produce "intelligent" results in complex tasks without human-level cognition. To employ a functional view of intelligence, such automated results can be considered "intelligent" to the extent that they approximate those that would have been produced by a similarly situated person employing high-level human cognitive processes. This is an outcome-oriented view of intelligence—assessing based upon

34. *See, e.g.*, David L. Schwartz, *Practice Makes Perfect? An Empirical Study of Claim Construction Reversal Rates in Patent Cases*, 107 MICH. L. REV. 223 (2008).

whether the results that were produced were sensible and useful—rather than whether the underlying process that produced them was "cognitive" in nature.

The machine learning spam filtering example illustrated this idea. We might normally think of the identification of spam email by a person as entailing a series of advanced cognitive processes. A human user determining whether a particular email is spam may do the following: visually process the email, read, absorb, and understand the language of the email text, contextualize the meaning of the email contents, reason about whether or not the email was solicited, and based upon that assessment determine whether the email constituted unwanted spam.[35]

One might conclude that, because spam determination involves intelligence when conducted by people, the task is inherently cognitive. In terms of automation, however, most of the advanced cognitive processes just described have not been artificially matched by computer systems to any significant degree.[36] Given that identifying spam emails appears to involve cognition, and that computers have not been able to replicate advanced human level cognitive processes—such as understanding arbitrary written text at the level of a literate person—one might presume it would not be possible to automate a task as abstract as identifying spam emails.[37]

However, in the example described earlier, the machine learning algorithm was able to automate the task of spam filtering through non-cognitive processes. Through the use of pattern detection, the algorithm was able to infer effective proxy markers for spam emails: that emails with the text "Earn Extra Cash" or with an origin from Belarus were statistically more likely to be spam. On that basis, the algorithm was able to make automated classifications that were useful and "intelligent" in

---

35. *See, e.g.*, Argye E. Hillis & Alfonso Caramazza, *The Reading Process and Its Disorders*, *in* COGNITIVE NEUROPSYCHOLOGY IN CLINICAL PRACTICE 229, 229–30 (David Ira Margolin ed., 1992) ("[A] cognitive process such as reading involves a series of transformations of mental representations. . . . On this view, even very simple cognitive tasks will involve various processing mechanisms . . . .").

36. RUSSELL & NORVIG, *supra* note 11, at 3–10.

37. For detailed explanations of the limits of Natural Language Processing (NLP) as of the writing of this Article, *see* RUSSELL & NORVIG, *supra* note 11, at 860–67; Robert Dale, *Classical Approaches to Natural Language Processing*, *in* HANDBOOK OF NATURAL LANGUAGE PROCESSING 1, 1–7 (Nitin Indurkhya & Frederick J. Damerau eds., 2d ed. 2010); Richard Socher et al., *Semantic Compositionality through Recursive Matrix-Vector Spaces*, *in* CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING § 1 (2012) (noting that particular NLP approaches are limited and "do not capture . . . the important quality of natural language that allows speakers to determine the meaning of a longer expression based on the meanings of its words and the rules used to combine them").

the sense that they approximated what a human user would have done after reading and comprehending the email.

However, notably, the algorithm did not engage in the meaning or substance of the email text in a manner comparable to a similarly situated person, nor did it need to.[38] In other words, the algorithm did not need to understand abstract concepts such as "email," "earning cash," "Belarus," or "spam"—in the way that a person does—in order to make accurate automatic spam classifications. Rather, it was able to detect statistical proxies for spam emails that allowed it to produce useful, accurate results, without engaging in the underlying meaning or substance of the email's constituent words.

Thus, the machine learning spam filter example illustrated a rather profound point: it is sometimes possible to accomplish a task typically associated with cognition not through artificial simulations of human intellectual processes, but through algorithms that employ heuristics and proxies—such as statistical correlations learned through analyzing patterns in data—that ultimately arrive at the same or similar results as would have been produced by a similarly situated intelligent person employing higher order cognitive processes and training.

1. Approximating Intelligence by Proxy

More generally, the example is illustrative of a broader strategy that has proven to be successful in automating a number of complex tasks: detecting proxies, patterns, or heuristics that reliably produce useful outcomes in complex tasks that, in humans, normally require intelligence.[39] For a certain subset of tasks, it may be possible to detect proxies or heuristics that closely track the underlying phenomenon without actually engaging in the full range of abstraction underlying that phenomenon, as in the way the machine learning algorithm was able to identify spam emails without having to fully understand substance and context of the email text. As will be discussed in Part III this is the principle that may allow the automation of certain abstract tasks within law that, when conducted by attorneys, require higher order cognition.

It is important to emphasize that such a proxy-based approach can have significant limitations. First, this strategy may only be appropriate for certain tasks for which approximations are suitable. By contrast, many complicated problems—particularly those that routinely confront attorneys—may not be amenable to such a heuristic-based technique.

---

38. SEGARAN, *supra* note 23, at 4.

39. *Id.* at 1–3.

For example, an attorney counseling a corporate client on a potential merger is a task of such scale, complexity, and nuance, with so many considerations, that a simple proxy approach would be inappropriate.

Second, a proxy-based strategy can often have significant accuracy limitations. Because proxies are stand-ins for some other underlying phenomenon, they necessarily are under- and over-inclusive relative to the phenomenon they are representing, and inevitably produce false positives and negatives. By employing proxies to analyze or classify text with substantive meaning for an abstract task, for example, such algorithms may produce more false positives or negatives than a similarly situated person employing cognitive processes, domain knowledge, and expertise. Thus, for example, automated spam-filters can often do a reasonably accurate job of classifying spam, but often make errors in substantively complex cases that would be trivial for a person to detect.[40] However, once the limitations are properly understood, for certain common purposes (e.g., classifying emails) where the efficiency of automation is more important than precision, such approximations may be sufficient.

2. *Developments in AI Research*

The strategy just described parallels changes among computer science artificial intelligence research over the last several decades. In the earliest era of AI research—from the 1950s through the 1980s—many researchers focused upon attempting to replicate computer-based versions of human cognitive processes.[41] Behind this focus was a belief that because humans employ many of the advanced brain processes to tackle complex and abstract problems, the way to have computers display artificial intelligence was to create artificial versions of brain functionality.[42]

However, more recently, researchers have achieved success in automating complex tasks by focusing not upon the intelligence of the automated processes themselves, but upon the results that automated processes produce.[43] Under this alternative view, if a computer system is able to produce outputs that people would consider to be accurate, appropriate, helpful, and useful, such results can be considered "intelligent"—even if they did not come about through artificial versions

---

40. YERAZUNIS, *supra* note 27, at 1–5.

41. *See, e.g.*, RUSSELL & NORVIG, *supra* note 11, at 3–10.

42. *Id.*

43. *See* Surden, *supra* note 5, at 685–86.

of human cognitive processes.

In general, this has been the approach followed by many successful AI systems of the past several years. These systems have used machine learning and other techniques to develop combinations of statistical models, heuristics, and sensors that would not be considered cognitive in nature (in that they do not replicate human-level cognition) but that produce results that are useful and accurate enough for the task required.[44] As described, these proxy-based approaches sometimes lack accuracy or have other limitations as compared to humans for certain complex or abstract tasks. But the key insight is that for many tasks, algorithmic approaches like machine learning may sometimes produce useful, automated approaches that are "good enough" for particular tasks.

A good example of this principle comes from the task of language translation. For many years, the translation of foreign languages was thought to be a task deeply connected with higher-order human cognitive processes.[45] Human translators of foreign languages call upon deep knowledge of languages, and abstract understanding of concepts, to translate foreign language documents. Many early AI projects sought to replicate in computers various language rules believed to reside within the human brain.[46] However, these early, bottom-up, rules-based language translation systems produced poor results on actual translations.[47]

More recent research projects have taken a different approach, using statistical machine learning and access to large amounts of data to produce surprisingly good translation results without attempting to replicate human-linguistic processes.[48] "Google Translate," for example, works in part by leveraging huge corpuses of documents that experts previously translated from one language to another. The United Nations (UN) has for instance, over the years, employed professional translators to carefully translate millions of UN documents into multiple languages, and this body of translated documents has become available in electronic

---

44. *See* RUSSELL & NORVIG, *supra* note 11, at 3–10.

45. *See* EKATERINA OVCHINNIKOVA, INTEGRATION OF WORLD KNOWLEDGE FOR NATURAL LANGUAGE UNDERSTANDING 215–20 (2012).

46. Mathias Winther Madsen, The Limits of Machine Translation 5–15 (Dec. 23, 2009) (unpublished Master thesis, University of Copenhagen), *available at* http://www.math.ku.dk/~m01mwm/The%20Limits%20of%20Machine%20Translation%20%28Dec.%2023,%202009%29.pdf.

47. *Id.*

48. *See* ENCYCLOPEDIA OF MACHINE LEARNING 912–13 (Claude Sammut & Geoffrey I. Webb eds., 2011).

form.[49] While these documents were originally created for other purposes, researchers have been able to harness this existing corpus of data to improve automated translation. Using statistical correlations and a huge body of carefully translated data, automated algorithms are able to create sophisticated statistical models about the likely meaning of phrases, and are able to produce automated translations that are quite good.[50] Importantly, the algorithms that produce the automated translations do not have any deep conception of the words that they are translating, nor are they programmed to understand the meaning and context of the language in the way a human translator might. Rather, these algorithms are able to use statistical proxies extracted from large amounts of previously translated documents to produce useful translations without actually engaging in the deeper substance of the language.

While this automated translation often falls short of expert human translations in terms of accuracy and nuance in many contexts, and may not be sufficient for tasks requiring high degrees of accuracy (e.g., translating legal contracts), the interesting point is that for many other purposes, the level of accuracy achieved by automated translation may be perfectly sufficient (e.g., getting a rough idea of the contents of a foreign web page).[51] Such automation has allowed for approximate but useful translations in many contexts where no translation was previously available at all.

In sum, the translation example illustrates a larger strategy that has proven successful in recent AI automation: applying machine learning analysis to large bodies of existing data in order to extract subtle but useful patterns that can be employed to automate certain complex tasks. Such pattern detection over large amounts of data can be used to create complex, nuanced computer models that can be brought to bear on problems that were previously intractable under earlier manual approaches to automation.

---

49. *See Find Out How Our Translations Are Created*, GOOGLE, http://translate.google.com/about (last visited Feb. 24, 2014).

50. *See id.*

51. *See* Madsen, *supra* note 46, at 10 (citing *Google Translate FAQ*, GOOGLE, http://www.google.com/-intl/en/help/faq_translation.html (last visited Mar. 25, 2009)).

III. MACHINE LEARNING AND LAW

A. *Machine Learning Applied to Law*

Because machine learning has been successfully employed in a number of complex areas previously thought to be exclusively in the domain of human intelligence, this question is posed: to what extent might these techniques be applied within the practice of law?[52] We have seen that machine learning algorithms are often able to build useful computer models of complex phenomena frequently by detecting patterns and inferring rules from data. More generally, we have seen that machine learning techniques have often been able to produce "intelligent" results in complex, abstract tasks, often not by engaging directly with the underlying conceptual substance of the information, but indirectly, by detecting proxies and patterns in data that lead to useful results. Using these principles, this Part suggests that there are a subset of legal tasks often performed manually today by attorneys, which are potentially partially automatable given techniques such as machine learning, provided the limitations are understood and accounted for.

I emphasize that these tasks may be *partially* automatable, because often the goal of such automation is not to replace an attorney, but rather, to act as a complement, for example in filtering likely irrelevant data to help make an attorney more efficient. Such a dynamic is discussed below in the case of automation in litigation discovery document review. There, the machine learning algorithms are not used to replace (nor are they currently capable of replacing) crucial attorney tasks such as of determining whether certain ambiguous documents are relevant under uncertain law, or will have significant strategic value in litigation. Rather, in many cases, the algorithms may be able to reliably filter out large swathes of documents that are likely to be irrelevant so that the attorney does not have to waste limited cognitive resources analyzing them. Additionally, these algorithms can highlight certain potentially relevant documents for increased attorney attention. In this sense, the algorithm does not replace the attorney but rather automates certain typical "easy-cases" so that the attorney's cognitive efforts and time can be conserved for those tasks likely to actually require higher-order legal skills.

There are particular tasks for which machine learning algorithms are

---

52. This is not to say that other AI techniques will not have an impact on the law. As I have written elsewhere, logic-based AI is impacting legal domains such as contracting. *See generally* Surden, *supra* note 5.

better suited than others. By generalizing about the type of tasks that machine learning algorithms perform particularly well, we can extrapolate about where such algorithms may be able to impact legal practice.

B. Predictive Models

1. Legal Predictions

Machine learning algorithms have been successfully used to generate predictive models of certain phenomena. Some of these predictive capabilities might be useful within the practice of law.[53]

The ability to make informed and useful predictions about potential legal outcomes and liability is one of the primary skills of lawyering.[54] Lawyers are routinely called upon to make predictions in a variety of legal settings. In a typical scenario, a client may provide the lawyer with a legal problem involving a complex set of facts and goals.[55] A lawyer might employ a combination of judgment, experience, and knowledge of the law to make reasoned predictions about the likelihood of outcomes on particular legal issues or on overall issue of liability, often in contexts of considerable legal and factual uncertainty.[56] On the basis of these predictions and other factors, the lawyer might counsel the client about recommended courses of action.

The ability to generally assess the likelihood of legal outcomes and relative levels of risk of liability in environments of considerable legal and factual uncertainty is one of the primary value-added functions of a good lawyer. As a general matter, attorneys produce such estimations by employing professional judgment, knowledge, experience, training, reasoning and utilizing other cognitive skills and intuitions.[57] However, as Daniel Katz has written, such prediction of likely legal outcomes may be increasingly subject to automated, computer-based analysis.[58] As

---

53. STEPHEN MARSLAND, MACHINE LEARNING: AN ALGORITHMIC PERSPECTIVE 103 (2011).

54. *See, e.g.*, Tanina Rostain, *Ethics Lost: Limitations of Current Approaches to Lawyer Regulation*, 71 S. CAL. L. REV. 1273, 1281–82 (1998); Brian Z. Tamanaha, *Understanding Legal Realism*, 87 TEX. L. REV. 731, 749–52 (2009).

55. *See, e.g.*, PAUL BREST & LINDA HAMILTON KRIEGER, PROBLEM SOLVING, DECISION MAKING AND PROFESSIONAL JUDGMENT 29–30 (2010).

56. *Id.*

57. *See, e.g.*, Patrick E. Longan, *The Shot Clock Comes to Trial: Time Limits for Federal Civil Trials*, 35 ARIZ. L. REV. 663, 687 (1993) ("Lawyers with trial experience and the consequent ability to predict outcomes more accurately can charge more.").

58. Katz, *supra* note 10, at 912.

Katz notes, there is existing data that can be harnessed to better predict outcomes in legal contexts.[59] Katz suggests that the combination of human intelligence and computer-based analytics will likely prove superior to that of human analysis alone, for a variety of legal prediction tasks.[60]

This Part will sketch a simple overview of what such an approach to legal prediction, involving machine learning, might look like. In general, such a method would involve using machine learning algorithms to automatically detect patterns in data concerning past legal scenarios that could then be extrapolated to predict outcomes in future legal scenarios. Through this process, an algorithm may be able to detect useful proxies or indicia of outcomes, and general probability ranges.

One relevant technique to apply to such a process is the "supervised learning" method discussed previously.[61] As mentioned, supervised learning involves inferring associations from data that has been previously categorized by humans.[62] Where might such a data set come from? Law firms often encounter cases of the same general type and might create such an analyzable data set concerning past cases from which associations could potentially be inferred. On the basis of information from past clients and combining other relevant information such as published case decisions, firms could use machine learning algorithms to build predictive models of topics such as the likelihood of overall liability. If such automated predictive models outperform standard lawyer predictions by even a few percentage points, they could be a valuable addition to the standard legal counseling approach. Thus, by analyzing multiple examples of past client data, a machine learning algorithm might be able to identify associations between different types of case information and the likelihood of particular outcomes.

For example, imagine that a law firm that represents plaintiffs in employment law cases records key data about past client scenarios into a database. Such data might include the nature of the incident, the type of company where the incident occurred, the nature of the claim. The firm could also keep track of the different aspects of the case, including the outcome of the case, whether it settled, how much it settled for, the judge involved, the laws involved, and whether it went to trial, etc. This data set of past case information that the firm has encountered over the

59. *Id.*

60. *Id.*

61. *See* FLACH, *supra* note 8, at 16–18.

62. *Id.*

years, combined with other data such as published case decisions or private sources of data about case outcomes, would be the "training set." And similar to the spam filter example, the machine learning algorithm could be trained to study the past examples to learn the salient features that are most indicative of future outcomes. Over time, after examining sufficient examples of past client cases, a machine learning algorithm could potentially build a predictive model determining the weights of the factors that are most predictive of particular outcomes.

For example, (to oversimplify) we could envision an algorithm learning that in workplace discrimination cases in which there is a racial epithet expressed in writing in an email, there is an early defendant settlement probability of 98 percent versus a 60 percent baseline. An attorney, upon encountering these same facts, might have a similar professional intuition that early settlement is likely given these powerful facts. However, to see the information supported by data may prove a helpful guide in providing professional advice.

More usefully, such an algorithm may identify a complex mix of factors in the data associated with particular outcomes that may be hard or impossible for an attorney to detect using typical legal analysis methods. For instance, imagine that the algorithm reveals that in cases in which there are multiple hostile emails sent to an employee, if the emails are sent within a three week time period, such cases tend to be 15 percent more likely to result in liability as compared to cases in which similar hostile emails are spread out over a longer one-year period. Such a nuance in timeframe may be hard for an attorney to casually detect across cases, but can be easily revealed through data pattern analysis. As such an algorithm received more and more exemplars from the training set, it could potentially refine its internal model, finding more such useful patterns that could improve the attorney's ability to make reasoned predictions.

In sum, entities concerned with legal outcomes could, in principle, leverage data from past client scenarios and other relevant public and private data to build machine learning predictive models about future likely outcomes on particular legal issues that could complement legal counseling. In essence, this would be formalizing statistically to some extent what lawyers often do intuitively today.[63] Lawyers who see

63. This is reminiscent of the quote from great mathematician Pierre-Simon Laplace who said several hundred years ago, "The theory of probabilities is at bottom nothing but common sense reduced to calculus; it enable us to appreciate with exactness that which accurate minds feels with a sort of instinct for which ofttimes they are unable to account." H. C. TIJMS, UNDERSTANDING PROBABILITY 3–4 (3d ed. 2012) (quoting LaPlace).

similar cases often over time develop an internal, intuitive understanding of the likely outcomes in particular cases once they factor in particular salient facts. Attorneys combine their judgment, training, reasoning, analysis, intuition, and cognition under the facts to make approximate legal predictions for their clients. To some extent, machine learning algorithms could perform a similar but complementary role, only more formally based upon analyzed data.

2. *Limitations to Machine Learning Legal Predictive Models*

There may be some limitations to predictive models that should be noted. Generally speaking, the goal of using machine learning is to analyze past data to develop rules that are generalizable going forward. In other words, the heuristics that an algorithm detects by analyzing *past* examples should be useful enough that they produce accurate results in *future*, never-before-seen scenarios. In the prior discussion for instance, the goal would be to analyze the data from past client scenarios, associate variables (e.g., hostile emails) with particular outcomes (e.g., increased settlement probability) in order to devise a set of heuristics that are sufficiently general that they would be predictive in cases with facts somewhat different from those in the training set. Such a learned model is thus only useful to the extent that the heuristics inferred from past cases can be extrapolated to predict novel cases.

There are some well-known problems with this type of generalization. First, a model will only be useful to the extent that the class of future cases have pertinent features in common with the prior analyzed cases in the training set.[64] In the event that future cases present unique or unusual facts compared to the past, such future distinct cases may be less predictable. In such a context, machine learning techniques may not be well suited to the job of prediction. For example, not every law firm will have a stream of cases that are sufficiently similar to one another such that past case data that has been catalogued contain elements that will be useful to predicting future outcomes. The degree of relatedness between future and past cases within a data-set is one important dimension to consider regarding the extent that machine learning predictive models will be helpful. Additionally, machine learning algorithms often require

64. There are other well-known problems with induction. Induction relies upon analyzing examples from the past to generalize about the future. However, under the so-called "Black Swan" problem, there may be never-before-seen, but salient scenarios that may arise in the future. In such an instance, a model trained upon past data may be insufficiently robust to handle rare or unforeseen future scenarios. *See, e.g.*, NASSIM NICHOLAS TALEB, THE BLACK SWAN: THE IMPACT OF THE HIGHLY IMPROBABLE 1–10 (2d ed. 2010).

a relatively large sample of past examples before robust generalizations can be inferred. To the extent that the number of examples (e.g., past case data) are too few, such an algorithm may not be able to detect patterns that are reliable predictors.

Another common problem involves overgeneralization. This is essentially the same problem known elsewhere in statistics as overfitting.[65] The general idea is that it is undesirable for a machine learning algorithm to detect patterns in the training data that are so finely tuned to the idiosyncrasies or biases in the training set such that they are not predictive of future, novel scenarios. For example, returning to the spam filter example, imagine the emails that were used as a training set happen to be systematically biased in some way: they all were sent from a data server located in Belarus. A machine learning algorithm may incorrectly infer from this biased training data that spam emails only originate from Belarus, and might incorrectly ignore spam emails from other countries. Such an inference would be accurate based upon the particular training data used, but as applied in the wider world, would produce inaccurate results because the training data was non-representative of spam emails generally.

Similarly, in the legal prediction context, the past case data upon which a machine learning algorithm is trained may be systematically biased in a way that leads to inaccurate results in future legal cases. The concern, in other words, would be relying upon an algorithm that is too attuned to the idiosyncrasies of the past case data that is being used to train a legal prediction algorithm. The algorithm may be able to detect patterns and infer rules from this training set data (e.g., examining an individual law firm's past cases), but the rules inferred may not be useful for predictive purposes, if the data from which the patterns were detected were biased in some way and not actually reflective enough of the diversity of future cases likely to appear in the real world.

A final issue worth mentioning involves capturing information in data. In general, machine learning algorithms are only as good as the data that they are given to analyze. These algorithms build internal statistical models based upon the data provided. However, in many instances in legal prediction there may be subtle factors that are highly relevant to legal prediction and that attorneys routinely employ in their professional assessments, but which may be difficult to capture in formal, analyzable data.

For example, imagine that there is an administrative board that

65. *See* RUSSELL & NORVIG, *supra* note 11, at 705.

adjudicates disciplinary cases and there has recently been a change in the board's personnel. An experienced attorney who has worked in a particular area for many years may be familiar with the board personnel and the types of cases that these individuals are and are not sympathetic to. Thus, such an attorney may make a recommendation as to a course of action to a client based upon a nuanced understanding of the board personnel and their particular inclinations. This might be the kind of information that would be available to an experienced attorney, and which is often used in legal counseling, but might be difficult to consistently and accurately capture in a data model. Consequently, a data model that does not include such hard-to-capture but predictive information may in fact produce inferior predictive results to an attorney.

Similarly, there are certain legal issues whose outcomes may turn on analyzing abstractions—such as understanding the overall public policy of a law and how it applies to a set of facts—for which there may not be any suitable data proxy. Thus, in general, if there are certain types of salient information that are both difficult to quantify in data, and whose assessment requires nuanced analysis, such important considerations may be beyond the reach of current machine learning predictive techniques.

C. Finding Hidden Relationships in Data

Machine learning techniques are also useful for discovering hidden relationships in existing data that may otherwise be difficult to detect. Using the earlier example, attorneys could potentially use machine learning to highlight useful unknown information that exists within their current data but which is obscured due to complexity. For example, consider a law firm that tracks client and outcome data in tort cases over the span of several years. A machine learning algorithm might detect subtle but important correlations that might go unnoticed through typical attorney analysis of case information. Imagine, for instance, that the algorithm detects that the probability of an early settlement is meaningfully higher when the defendant sued in a personal injury case is a hospital as compared to other types of defendants. This is the type of relationship that a machine learning algorithm might detect, and which may be relevant to legal practice, but might be subtle enough that it might escape notice absent data analysis.

In general, the mining of the law firm's existing data may give attorneys new information about important factors affecting outcomes (such as the category of the defendant as a hospital) that may otherwise escape traditional professional analysis. This represents a departure from

the normal mode of legal assessment of information. Attorneys typically rely upon internal intuition and previous experience to determine the factors that tend to be relevant to particular outcomes in particular instances. Machine learning as a technique—since it excels at ferreting out correlations—may help to supplement the attorney intuitions and highlight salient factors that might otherwise escape notice. The discovery of such embedded information, combined with traditional attorney analysis, could potentially impact and improve the actual advice given to clients.

1. Judicial Decisions and Data Relationships

There are some other potentially profound applications of machine learning models that can reveal non-obvious relationships, particularly in the analysis of legal opinions. A basis of the United States common law system is that judges are generally required to explain their decisions. Judges often issue major legal judgments in written opinions and orders.[66] In such a written document, judges typically explain why they decided the way that they did by referencing the law, facts, public policy, and other considerations upon which the outcome was based.[67]

Implicit in such a system of written opinions is the following premise: that the judge actually reached the outcome that she did for the reasons stated in the opinion. In other words, the justifications that a judge explicitly expresses in a written opinion should generally correspond to that judge's actual motivations for reaching a given outcome. Correspondingly, written legal decisions should not commonly and primarily occur for reasons other than those that were expressly stated and articulated to the public. At least one reason why legal opinions that do not reflect actual judicial motivations are undesirable is that there are thought to be certain motivations that are thought to be improper, illegal, or unseemly. For example, legal decisions based upon racial animus are illegal, and legal outcomes driven by pure partisanship over substance may be perceived as unseemly or improper. Moreover, it is desirable that stated judicial rationales correspond with actual rationales, because in a common law system, societal actors (and lawyers) rely upon legal opinions, and the stated justifications for these decisions, to make predictions about future legal outcomes and to understand and comply with the law.

---

66. Jonathan R. Macey, *Promoting Public-Regarding Legislation Through Statutory Interpretation: An Interest Group Model*, 86 COLUM. L. REV. 223, 253–54 (1986).

67. *Id.*

Since machine learning algorithms can be very good at detecting hard to observe relationships between data, it may be possible to detect obscured associations between certain variables in legal cases and particular legal outcomes. It would be a profound result if machine learning brought forth evidence suggesting that judges were commonly basing their decisions upon considerations other than their stated rationales. Dynamically analyzed data could call into question whether certain legal outcomes were driven by factors different from those that were expressed in the language of an opinion.

An earlier research project illustrated a related point. In that project, Theodore Ruger, Andrew Martin, and collaborators built a statistical model of Supreme Court outcomes based upon various factors including the political orientation of the lower opinion (i.e. liberal or conservative) and the circuit of origin of the appeal.[68] Not only did the statistical model outperform several experts in terms of predicting Supreme Court outcomes, it also highlighted relationships in the underlying data that may not have been fully understood previously.[69]

For example, the Supreme Court hears appellate cases originating from many different appellate circuits. Many experts had deemed the circuit of origin (e.g., Ninth or Sixth Circuits) of such a lower opinion as less important than other factors (e.g., the substantive law of the case) in relating to particular outcomes. However, the analysis of the data showed a stronger correlation between the circuit of origin and the outcome than most experts had expected based upon their intuition and judgment.[70] Although this earlier project did not involve machine learning algorithms in particular, it did involve some similar statistical techniques that might be used in a machine learning approach.

That project illustrates a basic point: that statistically analyzing decisions might bring to light correlations that could undermine basic assumptions within the legal system. If, for example, data analysis highlights that the opinions are highly correlated with a factor unrelated to the reasons articulated in the written opinions, it might lessen the legitimacy of stated opinions.[71] It also demonstrates the more general

---

68. Andrew D. Martin et al., *Competing Approaches to Predicting Supreme Court Decision Making*, 2 PERSP. ON POL. 761, 761–68 (2004); *see also* Theodore W. Ruger et al., *The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decision-Making*, 104 COLUM. L. REV. 1150, 1151–59 (2004).

69. Martin, *supra* note 68, at 761–68.

70. *Id.*

71. To be clear, this is not to suggest that correlation implies causation. It is perfectly consistent for Supreme Court decisions to be correlated with a non-substantive factor (e.g. circuit of origin) and still be based upon substantive determination. Thus, for example, if one circuit court was

point that statistical heuristics can be predictive and informative in a domain as abstract and full of uncertainty as law, even when computers do not actually engage with the underlying legal substance (e.g., underlying meaning and goals of the laws, doctrines, or policies) that is typically the primary focus of attorneys.

D. Document Classification and Clustering

The practice of law is intertwined with the production, analysis, and organization of text documents. These include written legal opinions, discovery documents, contracts, briefs, and many other types of written legal papers. Outside of law, machine learning algorithms have proven useful in automatically organizing, grouping, and analyzing documents for a number of tasks.[72] This Subpart will explore two machine learning methods that may be relevant to the automated analysis and organization of legal documents: 1) document classification; and 2) document clustering.

1. Automated Document Classification

In a document classification task, the goal of a machine learning algorithm is to automatically sort a given document into a particular, pre-defined category.[73] Often such classification is based upon the document's text and other document features.[74]

The earlier spam filtering example illustrated the idea of such an automated document classification. We can think of the machine learning algorithm described as attempting to classify a given incoming email document into one of two categories: unwanted spam or wanted email. The algorithm was able to make such automatic classifications based upon the various indicia of spam emails that it had automatically detected from past examples of spam (e.g., text included "Earn Extra Cash" or country of origin was Belarus). Moreover, the algorithm was able to "learn"—refine its internal model of the characteristics of spam emails as it examined more examples of spam—and improve in its classification ability over time as its internal model and rule-set of spam

---

consistently making errors in its interpretation of the law, one outcome (reversed) might be highly correlated with a particular circuit, but that outcome would not necessarily mean that the decision was being made based upon considering the circuit of origin.

72. SEGARAN, *supra* note 23, at 6–9.

73. *See, e.g.*, Kevin D. Ashley & Stefanie Brüninghaus, *Automatically Classifying Case Texts and Predicting Outcomes*, 17 ARTIFICIAL INTELLIGENCE & L. 125, 125–65 (2009).

74. *Id.*

became more sophisticated. Thus, we consider such a task to be "classification" because a human user, examining an email, is essentially performing the same classification task—deciding whether a particular incoming email is or is not in the category "spam."

Within law, there are numerous similar tasks that can be thought of as document classification problems. For these, machine learning algorithms may be useful, and in some cases have already been deployed.

2. *Classification of Litigation Docket Documents*

Since about 2002, documents associated with lawsuits have been typically contained in online, electronically accessible websites such as the Federal "PACER" court records system.[75] Such core documents associated with a lawsuit might include the complaint, multiple party motions and briefs, and the orders and judgments issued by the court. In a complicated court case, there may be several hundred documents associated with the case. However, obscured within such collections of hundreds litigation docket documents, there may be a few especially important documents—such as the active, amended complaint—that might be crucial to access, but difficult to locate manually. Electronic court dockets can become very lengthy, up to several hundred entries long. A particular important document—such as the active, amended complaint—may be located, for example, at entry 146 out of 300. Finding such an important document within a larger collection of less important docket entries often can be difficult.

The task of finding and organizing core case documents can be thought of as a document classification task. Analogous to the spam filtering example, a machine learning algorithm may be trained to learn the telltale characteristics that indicate that a particular document is a complaint rather than, say, a party motion. Such an algorithm could be trained to automate classifications of the documents based upon features such as the document text and other meta information such as the descriptive comments from the clerk of the court. Thus, key electronic court documents could be automatically identified as "complaints," "motions," or "orders," by machine learning algorithms, and parties could more easily to locate important docket documents thanks to such

---

75. *See* Administrative Office of the U.S. Courts, *25 Years Later, PACER, Electronic Filing Continue to Change Courts*, THE THIRD BRANCH NEWS (Dec. 9, 2013), http://news.uscourts.gov/25-years-later-pacer-electronic-filing-continue-change-courts; Amanda Conley et al., *Sustaining Privacy and Open Justice in the Transition to Online Court Records: A Multidisciplinary Inquiry*, 71 MD. L. REV. 772 (2012).

automated classification.

Projects such as the Stanford Intellectual Property Litigation Clearinghouse have employed similar machine learning techniques in order to automate the organization of very lengthy and complex case dockets, and to ease the finding of crucial court documents.[76] More broadly, machine learning algorithms are capable of providing intelligent classification of documents to aid in overall organization.

3. E-Discovery and Document Classification

Similarly, certain aspects of litigation discovery can be thought of as a document classification problem. In litigation discovery, each party is often presented with a voluminous trove of documents, including emails, memos, and other internal documents that may be relevant to the law and the facts at hand. A crucial task is sorting through such discovery documents in order to find those few that are actually relevant to some issue at hand. Thus, for example, in a case involving securities fraud, certain crucial emails demonstrating the intent to defraud may be extremely crucial to proving an element of the law. The major problem is that in modern litigation, the number of documents presented during discovery can be enormous, ranging from the tens of thousands to the millions.

Only an extremely small fraction of these documents are likely to be relevant to the issue or case at hand. In some sense, the task is akin to finding the proverbial needle (e.g., smoking-gun email) in the haystack (e.g., trove of millions of discovery documents). This task can be thought of as a classification task, as the goal is to classify each of the documents into a few categories based upon relevance, such as (for simplicity's sake), highly relevant, possibly relevant, likely irrelevant, highly irrelevant.

Previously, much of this discovery was conducted manually by junior associates who pored over and read emails and used their judgment to classify emails and other documents as either likely relevant or non-relevant.[77] In essence, this is similar to the classification task described above. The major difference is that the classification of an email as spam or not spam is often a dichotomous, binary classification—an email either is or is not spam. By contrast, the classification of a given

---

76. *Stanford IP Litigation Clearinghouse*, STAN. L. SCH., http://www.law.stanford.edu/organizations/programs-and-centers/stanford-ip-litigation-clearinghouse (last visited Jan. 27, 2014).

77. *See, e.g.*, John Markoff, *Armies of Expensive Lawyers, Replaced by Cheaper Software*, N. Y. TIMES (Mar. 4, 2011), http://www.nytimes.com/2011/03/05/science/05legal.html.

litigation discovery document as either relevant or non-relevant often exists upon a continuum of judgment. Some documents may be somewhat relevant, others highly relevant, and some not relevant at all. It is in this latter category that automation has proven highly useful.[78]

Today, certain aspects of litigation discovery are being automated in part, often by machine learning algorithms. Similar to the categorization tasks discussed before, in some cases, algorithms can roughly categorize documents by likelihood of relevance (often referred to as "predictive coding" or "technology assisted review"). In particular, they may be able to filter out documents that are likely irrelevant based upon dates or the parties involved. For example, such an algorithm may infer that emails that predated the core incident in the lawsuit by two years are highly likely to be irrelevant. There are, however, limitations to what these automated techniques can do. As discussed, the algorithms are not well suited to, or intended to, apply legal judgment in nuanced, uncertain areas. Rather, in many cases, the algorithms perform the role of filtering down the size of the document stack that is ultimately in need of lawyerly review. Once flagged, many of the documents still require attorney attention in order to conduct legal analysis as to relevance or privilege.

4. Clustering and Grouping of Related Documents

In a previous example, the machine learning algorithm described was used to classify documents into well-understood, predefined categories, such as "complaints," "motions," or "orders." In some cases, however, documents may have features in common, but the uniting characteristics of the documents may be unknown or non-obvious. In such an instance where there are hidden or unknown commonalities among items such as documents, a machine learning approach known as "clustering" may be useful.[79]

In clustering, a machine learning algorithm attempts to automatically group items that are similar in some way on the basis of some common characteristic that the algorithm has detected. In other words, the algorithm attempts to automatically detect hidden or non-obvious relationships between documents that would not otherwise be easily discoverable, and group such related documents together.

---

78. *See, e.g.*, Vincent Syracuse et al., *E-Discovery: Effects of Automated Technologies on Electronic Document Preservation and Review Obligations*, INSIDE COUNSEL (Dec. 18, 2012), http://m.insidecounsel.com/2012/12/18/e-discovery-effects-of-automated-technologies-on-e.

79. *See* RUI XU & DON WUNSCH, CLUSTERING 2–6 (2008).

In this way, such a machine algorithm might be used to discover that seemingly unconnected legal documents are actually related to one another in essential or useful ways. For example, imagine that there are two legal opinions in two fundamentally different areas of law: family law and trademark law. Imagine further that the two opinions share some subtle underlying commonality, such a lengthy discussion of best-practice strategies in administrative law. Such a connection between these two cases may go undetected by attorneys, since practitioners of family law may be unlikely to read trademark law opinions, and vice-versa. However, a clustering algorithm may be able to automatically find such an association and group the documents through this non-obvious relationship, by detecting a pattern among a large set of data—all opinions.

Consider another example in which automated document clustering and grouping might have uses within law. In patent law, patent examiners and patent attorneys spend a great deal of effort trying to find published documents describing inventions that are similar to a given patent.[80] Patent law has a requirement, for example, that the patent office not issue a patent on a patent application if the claimed invention is not new.[81] The way that one determines that an applied-for invention is not new is by finding "prior art" documents, which are documents that describe the invention but predate the patent application. Such prior art typically consists of earlier published scientific journal articles, patents, or patent applications that indicate than the invention had been created previously.

Given the huge volumes of published patents and scientific journals, it is a difficult task to find those particular prior art documents in the wider world that would prove that an invention was invented earlier. The task of finding such a document is essentially a problem involving automatically determining a relationship between the patent application and the earlier prior art document. Machine learning document clustering may potentially be used to help make the search for related prior art documents more automated and efficient by grouping documents that are related to the patent application at hand. More generally, automated document clustering might be useful in other areas of the law in which finding relevant documents among large collections is crucial.

---

80. JANICE M. MUELLER, PATENT LAW 30–40 (4th ed. 2012).

81. 35 U.S.C. § 102(a) (2006 & Supp. V 2011).

CONCLUSION

This Article focused upon a computer science approach known as machine learning and its potential impact upon legal practice. There has been a general view that because current AI technology cannot match the abstract analysis and higher-order cognitive abilities routinely displayed by trained attorneys, current AI techniques may have little impact upon law, barring significant technological advances. However, this Article has argued that outside of law, AI techniques—particularly machine learning—have been successfully applied to problems that had been traditionally thought to require human cognition.

This Article suggested that similarly, there are a number of tasks within the law for which the statistical assessments within the ambit of current machine learning techniques are likely to be impactful despite the inability to technologically replicate the higher-order cognition traditionally called upon by attorneys. The general insight is that statistical and other heuristic-based automated assessments of data can sometimes produce automated results in complex tasks that, while potentially less accurate than results produced by human cognitive processes, can actually be sufficiently accurate for certain purposes that do not demand extremely high levels of precision and accuracy.

DIGITAL DIRECTION FOR THE ANALOG ATTORNEY—DATA PROTECTION, E-DISCOVERY, AND THE ETHICS OF TECHNOLOGICAL COMPETENCE IN TODAY'S WORLD OF TOMORROW

Stacey Blaustein,[*] Melinda L. McLellan,[**] and James A. Sherer[***]

Cite as: Stacey Blaustein et al., *Digital Direction for the Analog Attorney—Data Protection, E-Discovery, and the Ethics of Technological Competence in Today's World of Tomorrow*, 22 RICH. J.L. & TECH. 10 (2016), http://jolt.richmond.edu/v22i4/article10.pdf.

I. INTRODUCTION

[1] Over the past twenty years, the near-constant use of sophisticated technological tools has become an essential and indispensable aspect of the practice of law. The time and cost efficiencies generated by these resources are obvious, and have been for years.[1] And because clients expect their counsel to take full advantage,[2] savvy attorneys understand that they must keep up with ever-evolving legal technologies to stay

[*] Stacey Blaustein is a Senior Attorney - Corporate Litigation with the IBM Corporation.
[**] Melinda L. McLellan is Counsel in the New York office of Baker & Hostetler LLP.
[***] James Sherer is Counsel in the New York office of Baker & Hostetler LLP.

[1] *See* Roger V. Skalbeck, *Computing Efficiencies, Computing Proficiencies and Advanced Legal Technologies*, VIRGINIA STATE BAR – RESEARCH RECOURSES (Oct. 2001), http://www.vsb.org/docs/valawyermagazine/oct01skalbeck.pdf, *archived at* https://perma.cc/8YWX-YAHF.

[2] *See* Ed Finkel, *Technology No Longer a 'Nice to Learn' for Attorneys*, LEGAL MANAGEMENT, ASSOCIATION OF LEGAL ADMINISTRATORS (Oct. 2014), http://encoretech.com/wp-content/uploads/2014/10/Technology-No-Longer-a-Nice-to-Learn-for-Attorneys_ALA-Legal-Management_Oct2014.pdf, *archived at* https://perma.cc/HUT3-672F.

competitive in a crowded marketplace.[3]

[2] With increased globalization and exponential growth in the creation, collection, use, and retention of electronic data, the challenges to all lawyers—especially those who may not have tech backgrounds or a natural aptitude for the mechanics of these innovations—are multiplying with breathtaking speed.[4] Nevertheless, many attorneys are either blissfully unaware of the power and potential danger associated with the tools they now find themselves using on a daily basis, or they are willfully avoiding a confrontation with reality. For lawyers, technological know-how is no longer a "nice to have" bonus; it now poses an ethical obligation. Where competent client representation demands a minimum level of tech proficiency, however, many lawyers come up short with respect to this fundamental component of their professional responsibilities.[5]

[3] What types of privacy and data security threats do various technologies pose to attorneys, their firms, their clients, and the legal

[3] *See, e.g.*, Evan Weinberger, *Fintech Boom Prompts Lawyers to Add Tech Know-How*, LAW360 (Sep. 4, 2015, 6:05 PM), http://www.law360.com/articles/692081/fintech-boom-prompts-lawyers-to-add-tech-know-how, *archived at* https://perma.cc/WVE8-UPGP; *see also* Allison O. Van Laningham, *Navigating in the Brave New World of E-Discovery: Ethics, Sanctions and Spoliation*, FDCC Q. 327(Summer 2007), http://www.thefederation.org/documents/V57N4-VanLaningham.pdf, *archived at* https://perma.cc/9L48-MPLU.

[4] *See* Frank Strong, *Beautiful Minds: 41 Legal Industry Predictions for 2016*, LEXISNEXIS LAWBLOG (Dec. 17, 2015), http://businessoflawblog.com/2015/12/legal-industry-predictions-2016/, *archived at* http://perma.cc/BG5W-R4DB.

[5] To further complicate matters, for attorneys and law firms practicing in the financial technology area such as payment, online lending, bitcoin and other virtual currencies, these lawyers need to be competent in "fintech", financial technology, another outgrowth of the expertise in technology requirement. *See* Evan Weinberger, *Fintech Boom Prompts Lawyers to Add Tech Know-How*, LAW360 (Sep. 4, 2015, 6:05 PM), http://www.law360.com/articles/692081/fintech-boom-prompts-lawyers-to-add-tech-know-how, *archived at* https://perma.cc/L76C-FZRL.

profession in general? What rules and regulations govern how attorneys may make use of technology in their practice, and how might clients seek to impose restrictions around such use when it comes to their corporate data? Must attorneys gain mastery over the intricate mechanics of the technological resources they employ, or is basic knowledge sufficient? How can we weigh the potential risks and rewards of cutting-edge, emerging digital products and electronic resources about which clients—and indeed, even the lawyers themselves—may understand very little? These are just a few of the questions that arise when we consider the issue of technological competence in the legal profession and corresponding ethical requirements.

[4] To begin to answer these questions, we look to the applicable Model Rules issued by the American Bar Association ("ABA"), various state-level professional ethics rules that incorporate the Model Rules, associated ethics opinions and guidance issued by the states, state and federal court decisions, and guidelines issued by sector-specific agencies and organizations.[6] Our focus in this investigation concerning lawyerly "technological competence" will be on privacy and data security risks and safeguards, e-Discovery-related challenges, and the potential perils of various uses of social media in the legal sphere.

II. THE THREAT LANDSCAPE: LAW FIRMS AS PRIME TARGETS

[5] In recent years, the volume and severity of attacks on electronically-stored data, and the information systems and networks that house that data, have increased exponentially. The modern-day "threat environment" is "highly sophisticated," and "massive data breaches are occurring with alarming frequency."[7] For attorneys, such perils implicate

[6] *See infra* Part III (explaining that agencies such as the FDA have issued guidance in their arena- Postmarket Management of Cybersecurity in Medical Devices).

[7] Report to the House of Delegates, ABA Cybersecurity Legal Task Force Section of Sci. & Tech. Law 1, http://www.americanbar.org/content/dam/aba/administrative/house_of_delegates/resoluti

multiple ethical and professional responsibilities with respect to how they handle data, including the duty to protect the confidentiality of client information and the obligation to provide "competent" representation.

[6] Unfortunately, law firms can provide a proverbial back door for hackers seeking access to a company's data, as attorneys often are custodians of a veritable "treasure trove" of valuable client information "that is extremely attractive to criminals, foreign governments, adversaries and intelligence entities."[8] Some hackers even focus their efforts primarily on law firms, especially those firms collecting vast amounts of data from corporate clients in the course of E-Discovery or corporate due diligence.[9] Corporate secrets, business strategies, and intellectual property all may be found in a law firm's collection of its clients' data.[10] In some cases, the interceptors may be looking for competitive information relevant to merger negotiations, or trying to suss out evidence of as-yet unannounced deals for insider trading purposes.[11]

[7] A 2015 report estimated that 80% of the biggest 100 law firms

---

ons/2014_hod_annual_meeting_109.authcheckdam.pdf, *archived at* https://perma.cc/KQT3-AFAJ.

[8] Ellen Rosen, *Most Big Firms Have Had Some Hacking: Business of Law*, BLOOMBERG (Mar. 11, 2015, 12:01 AM), http://www.bloomberg.com/news/articles/2015-03-11/most-big-firms-have-had-some-form-of-hacking-business-of-law, *archived at* https://perma.cc/YDR6-ZUV8.

[9] *See* Melissa Maleske, *A Soft Target for Hacks, Law Firms Must Step Up Data Security*, LAW360 (Sep. 23, 2015, 10:09 PM), http://www.law360.com/articles/706312/a-soft-target-for-hacks-law-firms-must-step-up-data-security, *archived at* https://perma.cc/6V7K-2WB4.

[10] *See id.*

[11] *See* Susan Hansen, *Cyber Attacks Upend Attorney-Client Privilege*, BLOOMBERG BUSINESSWEEK (Mar. 19, 2015, 2:56 PM), http://www.bloomberg.com/news/articles/2015-03-19/cyber-attacks-force-law-firms-to-improve-data-security, *archived at* https://perma.cc/29A5-MUNG.

have experienced some sort of data security incident.[12] And as is the case with so many companies that suffer a breach, law firms that *have* been hacked may not know about it for a considerable period of time. Moreover, unlike other industry sectors subject to various reporting requirements, law firms generally do not have a statutory obligation to publicly report cybercrimes that do not involve personally identifiable information.[13] Lack of obligations notwithstanding, a recent report indicated that "[t]he legal industry reported more "cyber threats" threats in January [2016] than nearly any other sector," topped only by the retail industry and financial services.[14]

[8] Although these reported "threats" might not necessarily result in data compromises, the fact that the legal industry frequently is among the most targeted for data theft should concern attorneys.[15] Anecdotal evidence of actual and attempted interference with law firms' data security systems abounds as well. In 2014, a report indicated that communications between lawyers from the law firm of Mayer Brown and officials with the Indonesian government were intercepted by an Australian intelligence agency that had ties with the U.S. National Security Agency ("NSA").[16] And the managing partner of the Washington-area offices of Hogan Lovells LLP recently noted that her firm "constantly intercept[s]

[12] *See* Rosen, *supra* note 8.

[13] *Id.*

[14] Mark Wolski, *Report: Legal Industry Was Heavily Targeted with Cyber Threats in January*, BLOOMBERG BNA (Mar. 9, 2016), https://bol.bna.com/report-legal-industry-was-heavily-targeted-with-cyber-threats-in-january, *archived at* https://perma.cc/ZCR9-2WRX.

[15] *See id.*

[16] James Risen & Laura Poitras, *Spying by N.S.A. Ally Entangled U.S. Law Firm*, N.Y. TIMES, Feb. 15, 2014, http://www.nytimes.com/2014/02/16/us/eavesdropping-ensnared-american-law-firm.html, *archived at* https://perma.cc/F8M4-TEQ7.

attacks."[17]

[9] The message to law firms seems clear: first, if "you're a major law firm, it's safe to say that you've either already been a victim, currently are a victim, or will be a victim."[18] Second, "[f]irms have to make sure they are not a weak link…which at its most basic level means their standards for protecting data need to be at least equivalent to those of the companies they represent."[19]

[10] It seems inevitable that client expectations and demands with regard to their legal service providers' security will continue to evolve and expand. One commentator recently predicted that in the future "clients across the board will demand firms demonstrate they're prepared for all shapes and sizes of cybersecurity breaches,"[20] while another prophesized that "in the name of risk management and data leakage prevention, a large financial industry corporation will challenge their outside counsel's [Bring Your Own Device] program."[21] Indeed, according to a 2014 report in the New York Times:

> Banks are pressing outside law firms to demonstrate that their computer systems are employing top-tier technologies to detect and deter attacks from hackers bent on getting their hands on corporate secrets for their own use of sale to others….Some financial institutions are asking law firms to fill out lengthy 60 page questionnaires detailing the [law

[17] *See* Rosen, *supra* note 8.

[18] *See* Hansen, *supra* note 11.

[19] Blake Edwards, *Verizon GC: Law Firms Prime Targets for Hackers*, BLOOMBERG BNA (Feb. 4, 2016), https://bol.bna.com/verizon-gc-law-firms-are-prime-targets-for-hackers/, *archived at* https://perma.cc/F6WU-N6FW.

[20] Strong, *supra* note 4.

[21] *Id.*

> firm's] cybersecurity measures, while others are demanding on-site inspections....Other companies are asking law firms to stop putting files on portable thumb drives, to stop emailing non-secure iPad or working on computers linked to a share network in countries like China and Russia.[22]

[11] In short, lawyers, law firms, and other legal services providers cannot afford to be complacent when it comes to cybersecurity.

A. Lawyering in the Cloud

[12] Firm adoption of cloud services is on the rise, especially among boutiques and solo practitioners that previously lacked the resources to compete effectively with larger law firms when it came to technology and data storage.[23] At first, the added value of cloud services created a perception that "nirvana had arrived" in terms of leveling the playing field for smaller firms.[24] Notwithstanding the apparent advantages of the cloud, attorneys were quick to identify concerns associated with the technology and its supporting practices, including "increased sensitivity to cyber-threats and data security."[25] Some commentators opted for a cautious and conservative approach, noting that the "legal profession has developed many safeguards to protect client confidences," and that the use of cloud hosting, among other practices, fell on a continuum where, as "an individual attorney gives up direct control of his or her client's

---

[22] Matthew Goldstein, *Law Firms Are Pressed on Security for Data*, N.Y. TIMES (Mar. 26, 2014), http://dealbook.nytimes.com/2014/03/26/law-firms-scrutinized-as-hacking-increases/, *archived at* https://perma.cc/Q77A-8BN3.

[23] *See* N.Y. CITY BAR COMM. ON SMALL LAW FIRMS, THE CLOUD AND THE SMALL LAW FIRM: BUSINESS, ETHICS AND PRIVILEGE CONSIDERATIONS 2 (Nov. 2013), http://www2.nycbar.org/pdf/report/uploads/20072378-TheCloudandtheSmallLawFirm.pdf, *archived at* https://perma.cc/A8EG-AH7E.

[24] *Id.*

[25] Strong, *supra* note 4.

information, he or she takes calculated risks with the security of that information."[26]

[13] There is hope for attorneys drawn to the advantages of cloud services, but vigilance and diligence is required. As noted in tech law guidance from March 2014, "[u]sing the cloud to hold data is fine, so long as you understand the security precautions."[27] Security concerns have put a damper on adoption rates and the development of attorney-specific cloud services lags behind other industries. This reluctance is unsurprising given the slow rate of technological advancements within the profession generally,[28] and a deserved reputation that the tendency of firms is "to be technology followers, not leaders."[29] That said, lawyers do seem to be embracing the cloud to some extent,[30] with the majority utilizing cloud

[26] Patrick Mohan & Steve Krause, *Up in the Cloud: Ethical Issues that Arise in the Age of Cloud Computing*, 8 ABI ETHICS COMM. NEWS L. 1 (Feb. 2011), http://www.davispolk.com/sites/default/files/files/Publication/a2e048ea-3b12-45fe-a639-9fc2881a4db8/Preview/PublicationAttachment/0f8af440-1db0-4936-8d0d-a1937a0e6c8f/skrause.ethics.clouds.feb11.pdf, *archived at* https://perma.cc/SW3C-FYT5.

[27] Sharon D. Nelson & John W. Simek, *Why Do Lawyers Resist Ethical Rules Requiring Competence with Technology?*, SLAW (Mar. 27, 2015), http://www.slaw.ca/2015/03/27/why-do-lawyers-resist-ethical-rules-requiring-competence-with-technology/, *archived at* https://perma.cc/6HNN-UCDZ.

[28] Ed Finkel, *Technology No Longer a 'Nice to Learn' for Attorneys*, Legal Management, Association of Legal Administrators (Oct. 2014) http://encoretech.com/wp-content/uploads/2014/10/Technology-No-Longer-a-Nice-to-Learn-for-Attorneys_ALA-Legal-Management_Oct2014.pdf, *archived at* https://perma.cc/TW7N-4WP5.

[29] Leslie Pappas, *The Security Concerns Holding Up One Firm's Cloud Usage*, BLOOMBERG BNA (Jan. 22, 2016), https://bol.bna.com/the-security-concerns-holding-up-one-firms-cloud-usage/, *archived at* https://perma.cc/Z4LJ-H83Q.

[30] *See* Casey C. Sullivan, *Is It Time for a Law Firm Cloud Computing Security Standard?*, FINDLAW (Feb. 18, 2016), http://blogs.findlaw.com/technologist/2016/02/is-it-time-for-a-law-firm-cloud-computing-security-standard.html, *archived at* https://perma.cc/78HF-KKX4.

solutions in some capacity,[31] even if implementation is mostly through "sporadic action and adoption among firms and law departments."[32]

[14] With respect to professional obligations, this type of implementation may not require specific technological expertise on the part of the attorneys. New York State Bar Association Opinion 1020, which addressed ethical implications of the "use of cloud storage for purposes of a transaction," determined that compliant usage "depends on whether the particular technology employed provides reasonable protection to confidential client information and, if not, whether the lawyer obtains informed consent from the client after advising the client of the relevant risks."[33]

[15] Further, New Jersey Opinion 701 addresses the reality that it is

> [N]ot necessarily the case that safeguards against unauthorized disclosure are inherently stronger when a law firm uses its own staff to maintain a server. Providing security on the Internet against hacking and other forms of unauthorized use has become a specialized and complex facet of the industry, and it is certainly possible that an independent [Internet Service Provider] may more efficiently and effectively implement such security precautions.[34]

---

[31] *See* Jonathan R. Tung, *Survey: Law Departments Are Warming Up to the Cloud*, FINDLAW (Feb. 18, 2016), http://blogs.findlaw.com/in_house/2016/02/survey-law-depts-are-warming-up-to-the-cloud.html, *available at* https://perma.cc/M89M-LC3M.

[32] Strong, *supra* note 4.

[33] N.Y. State Bar Ass'n Comm. on Prof'l Ethics, Op. 1020 (Sept. 12, 2014), http://www.nysba.org/CustomTemplates/Content.aspx?id=52001, *archived at* https://perma.cc/8MPU-62BR.

[34] N.J. Advisory Comm. on Prof'l Ethics, Op. 701 (2006), https://www.judiciary.state.nj.us/notices/ethics/ACPE_Opinion701_ElectronicStorage_12022005.pdf, *archived at* https://perma.cc/2H5Y-UYWX.

[16] Opinion 701 does include an additional caveat, that

> [W]hen client confidential information is entrusted in unprotected form, even temporarily, to someone outside the firm, it must be under a circumstance in which the outside party is aware of the lawyer's obligation of confidentiality, and is itself obligated, whether by contract, professional standards, or otherwise, to assist in preserving it.[35]

B. E-Discovery Tools

[17] To begin with, federal judges are unconvinced that many of the attorneys appearing before them understand how to make proper use of the technologies and related strategies associated with E-Discovery. A recent report, "Federal Judges Survey on E-Discovery Best Practices & Trends,"[36] compiled some of the judges' concerns, noting first "the typical attorney…does not have the legal and technical expertise to offer effective advice to clients on e-discovery."[37] Some of the judges' comments were quite blunt, with one noting that "[s]ome attorneys are highly competent; but most appear to have significant gaps in their understanding of e-discovery principles."[38]

[18] Legal ethical rules and related opinions and scholarship provide guidance for what attorney E-Discovery competence should look like. At least one author has made the connection between professional responsibility and technological savoir-faire, noting that:

---

[35] *Id.*

[37] Aebra Coe, *Judges Lack Faith in Attys' E-Discovery Skills, Survey Says*, LAW360 (Jan. 28, 2016), http://www.law360.com/articles/751961/judges-lack-faith-in-attys-e-discovery-skills-survey-says, *archived at* https://perma.cc/5UJB-D2YX.

[38] *Id.*

> There is growing recognition across the country that the practice of law requires some degree of competence in technology. In the forum of litigation, competence in technology necessarily equates with competence in e-discovery. It is only a matter of time before ethics bodies across the nation call for competence in e-discovery.[39]

[19] The opinions of courts and bar associations may carry the most weight, but a number of influential professional and industry groups also have offered useful commentary on technological competence. For example, competence is

> …highlighted in the very first rule of legal ethics, according to the American Bar Association['s] Rule 1.1 of the ABA Model Rules of Professional Conduct," which "specifically recognized the need for technological competence through a significant change in August 2012 that formally notified all lawyers (and specifically those in jurisdictions following the Model Rules) that competency includes current knowledge of the impact of e-Discovery and technology on litigation.[40]

[20] This guidance predated and perhaps presaged a number of state and federal reactions to technology and the impact of these developments on the practice of law, especially within the realm of E-Discovery. Delaware amended its Lawyers' Rules of Professional Conduct as they

---

[39] Bob Ambrogi, *California Considers Ethical Duty to Be Competent in E-Discovery*, CATALYST BLOG (Feb. 27, 2015), http://www.catalystsecure.com/blog/2015/02/california-considers-ethical-duty-to-be-competent-in-e-discovery/, *archived at* https://perma.cc/2FXD-8KM4.

[40] Karin S. Jenson, Coleman W. Watson & James A. Sherer, *Ethics, Technology, and Attorney Competence*, THE ADVANCED EDISCOVERY INST. (Nov. 2014), http://www.law.georgetown.edu/cle/materials/eDiscovery/2014/frimorndocs/EthicsIneDiscoveryBakerHostetler.pdf, *archived at* https://perma.cc/TFR6-VZNG.

related to technology in 2013;[41] North Carolina[42] and Pennsylvania[43] did the same shortly thereafter.

[21] California's relatively recent Formal Opinion No. 2015-193 (the "California Opinion") addresses a number of issues associated with attorney ethical duties vis-à-vis E-Discovery. Although advisory in nature, the California Opinion states "attorneys have a duty to maintain the skills necessary to integrate legal rules and procedures with 'ever-changing technology.'"[44] That reads broadly, but the California Opinion has been interpreted to indicate that, because E-Discovery arises "in almost every litigation matter, attorneys should have at least a baseline understanding of it."[45] Specifically, the California Opinion begins with the premise that E-Discovery requires an initial assessment of its inclusion at the beginning of a matter.[46] If E-Discovery will be a component of a matter,

> [T]he duty of competence requires an attorney to assess his or her own e-discovery skills and resources as part of the attorney's duty to provide the client with competent

[41] *See* Order Amending Rules 1.0, 1.1, 1.4, 1.6, 1.17, 1.18, 4.4, 5.3, 5.5, 7.1, 7.2, and 7.3 of the Delaware Lawyers' Rules of Professional Conduct, DEL. R. PROF'L CONDUCT (2013), http://courts.delaware.gov/rules/pdf/dlrpc2013rulechange.pdf.

[42] *See* N.C. STATE. BAR RULES OF PROF'L RESPONSIBILITY & CONDUCT R. 1.1 (2014), http://www.ncbar.com/rules/rules.asp?page=4, *archived at* https://perma.cc/7R44-4JAG.

[43] *See* Notice of Proposed Rulemaking, 43 Pa. Bull. 1997 (Apr. 13, 2013), http://www.pabulletin.com/secure/data/vol43/43-15/652.html, *archived at* https://perma.cc/WS5G-MHKQ.

[44] Bob Ambrogi, *California Finalizes Ethics Opinion Requiring Competence in E-Discovery*, CATALYST BLOG (Aug. 6, 2015), https://www.catalystsecure.com/blog/2015/08/california-finalizes-ethics-opinion-requiring-competence-in-e-discovery/, *archived at* https://perma.cc/V7NV-QCWW.

[45] *Id.*

[46] *See id.*

> representation. If an attorney lacks such skills and/or resources, the attorney must try to acquire sufficient learning and skill, or associate or consult with someone with expertise to assist.[47]

[22] Other commentators have noted that the California Opinion focuses on "nine (9) core competency issues" which would offer "solid guidelines for attorneys…to maintain competency and protect client confidentiality in the era of eDiscovery."[48] One author notes that one of these core competency issues and its related directive, that of performing data searches, stretches across the entirety of the E-Discovery process "occurring at each of these steps, from preservation and collection to review and redaction."[49]

[23] Soon after the California Opinion was decided, Magistrate Judge Mitchell Dembin issued a Southern District of California decision that addressed "counsels' ethical obligations and expected competency" in *HM Electronics, Inc. v. R.F. Technologies, Inc.*[50] The *HM Electronics* case focused both on specific steps the attorneys *should have taken* (such as

---

[47] State Bar of Cal. Standing Comm. on Prof'l Responsibility & Conduct, Formal Op. 2015-193 (2015), https://ethics.calbar.ca.gov/Portals/9/documents/Opinions/CAL%202015-193%20%5B11-0004%5D%20(06-30-15)%20-%20FINAL.pdf, *archived at* https://perma.cc/8GWJ-BVJ2.

[48] Adam Kuhn, *The California eDiscovery Ethics Opinion: 9 Steps to Competency*, RECOMMIND BLOG (Aug. 11, 2015), http://www.recommind.com/blog/california-ediscovery-ethics-opinion-9-steps-to-competency, *archived at* https://perma.cc/2X2K-FCRQ.

[49] *Id.*

[50] H. Christopher Boehning & Daniel J. Toal, *E-Discovery Competence of Counsel Criticized in Sanctions Decision*, NEW YORK LAW JOURNAL (Oct. 6, 2015), http://www.newyorklawjournal.com/id=1202738840840/EDiscovery-Competence-of-Counsel-Criticized-in-Sanctions-Decision#ixzz42wNK34Ms, *archived at* https://perma.cc/4BMP-T76U.

implementing a legal hold and doing the legwork necessary to certify discovery responses as true) as well as behavior actively detrimental to the case (instructing client personnel to destroy relevant documents).[51] Of note in Judge Dembin's excoriation of the misbehaving attorneys is his statement that "a judge must impose sanctions for a violation of the Rule that was without substantial justification."[52] One article suggests that part of the problem may be simply that "counsel and clients alike…fail to take seriously judges' expectations for how they conduct themselves throughout the discovery process."[53]

[24] New York attorneys followed the California Opinion with interest, first noting that it merely presented "the standard tasks one should engage in and competently execute to properly collect and produce responsive ESI [Electronically Stored Information] to the opposing party."[54] A 2009 S.D.N.Y. opinion had chastised attorneys who would otherwise disclaim experience, warning that it was "time that the Bar—even those lawyers who did not come of age in the computer era" understood E-Discovery technologies and their application.[55] A recent article indicated that there is "an ample basis to discern a framework for ethical obligations, derived from ethics rules, court rules, and sanctions decisions in the e-discovery

[51] *See generally* HM Elecs., Inc. v. R.F. Techs., Inc., 2015 U.S. Dist. LEXIS 104100 (S.D. Cal. Aug. 7, 2015) (arguing the invalidity of the steps that the defendants took in order to certify discovery as true).

[52] Boehning & Toal, *supra* n. 50.

[53] *Id.*

[54] Samantha V. Ettari & Noah Hertz-Bunzl, *Ethical E-Discovery: Core Competencies for New York Lawyers*, NEW YORK LAW JOURNAL (Nov. 2, 2015), http://www.kramerlevin.com/files/Publication/60607051-f018-43b7-8a3c-7d43b4ff6e50/Presentation/PublicationAttachment/1e570a52-e27d-425f-a75b-9e25811df796/NYLJ%20Article-EDiscovery%2011.2.15.pdf, *archived at* https://perma.cc/F3R8-UWM6.

[55] William A. Gross Constr. Assocs., Inc. v. Am. Mfrs. Mut. Ins. Co., 256 F.R.D. 134, 136 (S.D.N.Y. 2009).

context" based in part on the history of New York courts as "leaders in the advancement of e-discovery law."[56]

[25] But such a "framework for ethical obligations" might not even be necessary where competence is the ethical rule at issue. Competence "requires that lawyers have the legal knowledge, skill, thoroughness, and preparation to conduct the representation, or associate with a lawyer who has such skills"[57] and that supervision is appropriate to ensure that the work of others "is completed in a competent manner."[58] The issue of supervision came up in another advisory opinion, Ethics Opinion 362 of the District of Columbia Bar, which indicated that retaining an e-Discovery vendor that provided all of the E-Discovery services was both impermissible (as the unauthorized practice of law on the part of the vendor) as well as a circumstance where the attorney engaging such a vendor was not absolved from understanding and supervising the work performed, no matter how technical.[59]

1. Metadata in Electronic Files

[26] A very basic threat to client confidentiality (as well as the secrecy of counsel's strategy) is the existence of metadata embedded in electronic files exchanged between the parties or produced as evidence. Most frequently this threat exists in the form of automatically created information about a file, including changes made to the file, that can be recovered and viewed by a third party if not removed (or "scrubbed") prior

[56] *See* Ettari & Hertz-Bunzl, *supra* n. 54.

[57] *See* Ettari & Hertz-Bunzl, *supra* n. 54 (citing New York Rules of Professional Conduct (N.Y. Rule) 1.1.5).

[58] *See* Ettari & Hertz-Bunzl, *supra* n. 54 (citing N.Y. Rule 5.1(c)).

[59] *See generally* D.C. Comm. on Legal Ethics, Formal Op. 362 (2012), https://www.dcbar.org/bar-resources/legal-ethics/opinions/opinion362.cfm, *archived at* https://perma.cc/TXA5-26ZG (discussing the permissibility of non-lawyer ownership of discovery service vendors).

to disclosing the file. This "application metadata" can include information about the document itself, the author, comments and prior edits, and may also detail when the document was created, viewed, modified, saved or printed.[60] In addition to the fact that access to metadata can provide opposing parties with everything from revealing insights to damning evidence, there's also a "real danger" that "application metadata may be inaccurate."[61]

[27] Further, disputes related to metadata regularly arise in the E-Discovery context. Indeed, one of the "biggest challenges in electronic discovery" concerns "[u]nderstanding when metadata is relevant and needs to be preserved and produced."[62] To cite just one example, the concurring opinion in *State v. Ratcliff* noted that judges must determine whether submitted evidence contained more than the information visible on the face of the document, or whether metadata was included as well, where the distinction "is critical, both on an ethical and adjudicative basis."[63]

[28] Accordingly, understanding and managing metadata has become a baseline requirement for technological competence when dealing with client data and attorney work product. Numerous products exist to help save lawyers from themselves when it comes to accidental disclosure of metadata, including software applications that may be integrated into email programs to prevent documents from being sent outside the network

---

[60] *See generally* The Sedona Conference Working Group, *Best Practices Recommendations & Principles for Addressing Electronic Document Production*, THE SEDONA PRINCIPLES: SECOND EDITION, June 2007, at 60, 61 https://thesedonaconference.org/publication/The%20Sedona%20Principles, *archived at* https://perma.cc/UU5K-V8KQ (explaining the composition and functionality of metadata).

[61] *Id.* at 4.

[62] *Id.*

[63] State v. Ratcliff, 849 N.W.2d 183, 196 (N.D. 2014).

without first passing through a scrubbing filter. And the e-filing portal in many jurisdictions "contains a warning reminder that it is the responsibility of the e-filer to strip metadata from the electronic file before submitting it through the portal."[64] Reliance on these tools, however, may not suffice for long as the sophistication and complexity of issues related to the creation and manipulation of metadata continue to evolve.

III. OVERVIEW OF U.S. DATA PRIVACY AND INFORMATION SECURITY LAW

[29] The sectoral approach to privacy and data security law in the United States often is described as "a patchwork quilt" comprised of numerous state and federal laws and regulations that apply variously to certain types of data, certain industries, the application of particular technologies, or some combination of those elements. These laws may be enforced by a variety of regulators, with state Attorneys General and the Federal Trade Commission often leading the way.[65] Plaintiffs' lawyers also are prominent actors in this space, bringing an ever-increasing number of class action and other civil suits alleging violations of privacy rights, data protection laws, and information security standards.

[30] Although there are no federal or state privacy statutes specifically applicable solely to lawyers, numerous data protection laws and regulations may apply to attorneys in their role as service provider to their clients or in other contexts. The obligations associated with these laws often implicitly or explicitly demand that lawyers handling client data (1) have a thorough understanding of the potential privacy and security

[64] *See* Christian Dodd, *Metadata 101 for Lawyers: A 2-Minute Primer*, LAW360 (Oct. 15, 2015, 4:30 PM), http://www.law360.com/articles/712714/metadata-101-for-lawyers-a-2-minute-primer, *archived at* https://perma.cc/3VCT-TJRB.

[65] *See* Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583, 587 (2014).

risks to that data; (2) assess and determine how best to secure the data and prevent unauthorized access to the data; and (3) supervise anyone acting on their behalf with respect to the data to ensure the data is appropriately protected at all times.

[31] Below we describe a few of the privacy and data security laws that tend to come up frequently for lawyers and impose requirements on their handling of client data that may involve technological competence. This discussion is by no means exhaustive, as technology touches upon virtually every aspect of data protection regulation and information security counseling by attorneys in the field. To provide just a few examples, advising companies on restrictions applicable to cross-border data transfers, data localization requirements, cybersecurity standards and information sharing obligations, and regulatory action around the use of biometrics and geolocation technologies are just a few examples of areas where a lawyer must have an understanding of the underlying technology to effectively assist clients.

A. HIPAA – Business Associate Agreements

[32] The Health Insurance Portability and Accountability Act of 1996 ("HIPAA"), is the most significant health privacy law in the United States, imposing numerous obligations on "covered entities" and "business associates" of those "covered entities" to protect the privacy and security of "protected health information" ("PHI").[66] As required by HIPAA, the Department of Health and Human Services ("HHS") issued two key sets of regulations to implement the statute: the Privacy Rule[67] and the Security Rule.[68]

---

[66] *See* Health Insurance Portability and Accountability Act of 1996 (HIPAA), 42 U.S.C. §§1320d to 1320d-8 (2007) [hereinafter HIPAA].

[67] *See* Standards for Privacy of Individually Identifiable Health Information, 65 Fed. Reg. 82,462 (Dec. 28, 2000) (codified at 45 C.F.R. pts. 160, 164).

[68] *See* Security Standards, 68 Fed. Reg. 8333, 8334 (Feb. 20, 2003) (codified at 45 C.F.R. pts. 160, 162, 164).

[33] Although attorneys and law firms are not themselves considered covered entities directly subject to HIPAA's requirements,[69] when attorneys obtain PHI from covered entity clients in the course of a representation, the law firm may be subject to certain HIPAA Privacy Rule requirements[70] in its role as a business associate.[71] The Privacy Rule and the Security Rule apply to a covered entity's interactions with third parties (e.g., service providers) that handle PHI on the covered entity's behalf.[72] The covered entity's relationships with these "business associates" are governed by obligatory contracts known as business associate agreements ("BAAs") that must contain specific terms.[73] With respect to technological competence specifically, for example, the BAA requires the business associate to implement appropriate safeguards to prevent use or disclosure of PHI other than as provided for by the BAA, and states that the business associate must ensure that any agents/subcontractors that receive PHI from the business associate also protect the PHI in the same manner. And attorneys who "hold HIPAA data or [other PII] may be governed by state or federal law beyond the scope of the proposed rules, which is noted in the new comments"[74] to ABA Rule

[69] The health plan within an organization, such as a law firm's employee health plan, may itself be a "covered entity" for HIPAA compliance purposes, but a firm generally is not, itself, a covered entity. *See, e.g.*, HIPAA, *supra* note 66.

[70] *See* John V. Arnold, *PRIVACY: What Lawyers Must Do to Comply with HIPAA*, 50 TENN. B.J. 16, 17 (Mar. 2014).

[71] *See* Lisa J. Acevedo et. al., *New HIPAA Liability for Lawyers*, 30 GPSOLO, no. 4, 2013, http://www.americanbar.org/publications/gp_solo/2013/july_august/new_hipaa_liability_lawyers.html, *archived at* https://perma.cc/F88Y-U928.

[72] *See* Standards for Privacy of Individually Identifiable Health Information, *supra* note 67; *see* Security Standards, *supra* note 68.

[73] Both the Privacy Rule and the Security Rule dictate certain terms that must be included in a BAA.

[74] *See* Nelson & Simek, *supra* note 27.

1.6, discussed further below.

B. GLBA Safeguards Rule Requirements

[34] Pursuant to the Gramm-Leach-Bliley Act ("GLBA"), the primary federal financial privacy law in the United States, various federal agencies promulgated rules and regulations addressing privacy and data security issues.[75] For example, the Safeguards Rule requires financial institutions to protect security of personally identifiable financial information by maintaining reasonable administrative, technical, and physical safeguards for customer information.[76] To comply with the Safeguards Rule, a financial institution must develop, implement, and maintain a comprehensive information security program, and that program must address the financial institution's oversight of service providers that have access to customers' nonpublic personal information ("NPI").[77]

[35] Again, although a law firm is not a financial institution directly subject to the GLBA, when it acts as counsel to a financial institution, GLBA requirements may apply to its handling of NPI received from that client. To the extent a financial institution's law firm will have access to such NPI in the course of the representation, the financial institution-client must take reasonable steps to ensure the law firm has the ability to safeguard such data prior to disclosing it to the firm, and require the firm to contractually agree (in writing) to safeguard the NPI. Assuming such data will be stored electronically (a safe assumption in virtually all cases), it is incumbent on the law firm to understand the potential data security risks and how to prevent unauthorized access, use, transfer, or other processing of their clients' NPI.

---

[75] *See* 15 U.S.C. §§ 6801–6809 (2012).

[76] *See* 16 C.F.R. §§ 314.2, 314.3(b).

[77] *See* 16 C.F.R. § 314.4(a-c).

C. State Data Security Laws

[36] At the state level, there are numerous laws and regulations regarding the protection of personal information (and other types of data) that apply to all entities that maintain such data, including lawyers, law firms, and other legal service providers.

[37] A number of states, such as California, Connecticut, Maryland, Nevada, Oregon, and Texas, have enacted laws that require companies to implement information security measures to protect personal information of residents of the state that the business collects and maintains.[78] These laws of general application are relevant to attorneys and law firms with respect to the personal information they maintain—both client data and data relating to their employees. Typically, these laws are not overly prescriptive and include obligations to implement and maintain reasonable security policies and procedures to safeguard personal information from unauthorized access, use, modification, disclosure, or destruction (though most do not offer a definition or description of what is meant by "reasonable" security). Some laws, such as California's, impose a requirement to contractually obligate non-affiliated third parties that receive personal information from the business to maintain reasonable security procedures with respect to that data.[79]

[38] Massachusetts was the first state to enact regulations that directed businesses to develop and implement comprehensive, written information security programs ("WISPs") to protect the personal information of Massachusetts residents.[80] These regulations apply to all private entities

[78] *See, e.g.*, CAL. CIV. CODE § 1798.81.5 (Deering 2009); CONN. GEN. STAT. § 42-471 (2010); MD. CODE ANN., COM. LAW §§ 14-3501 to 14-3503 (LexisNexis 2009); NEV. REV. STAT. § 603A.210 (2009); OR. REV. STAT. § 646A.622 (2009); TEX. BUS. & COM. CODE ANN. §§ 72.001–72.051 (West 2009).

[79] *See* CAL. CIV. CODE § 1798.81.5 (Deering 2009).

[80] *See* 201 MASS. CODE REGS. 17.01–17.05 (2008).

(including law firms) that maintain personal information of Massachusetts residents, including those that do not operate in Massachusetts; they also list a number of minimum standards for the information security program.[81] The Massachusetts regulations are relatively prescriptive as compared to other similar state laws of this nature, and they include numerous specific technical requirements.

[39] These requirements apply to law firms directly, but they also apply to law firms as service providers to businesses that maintain personal information of Massachusetts residents. A compliant WISP must address the vetting of service providers, and the contract must include provisions obligating the service provider to protect the data.[82]

IV. APPLICABLE ETHICAL RULES AND GUIDANCE

[40] The myth of the Luddite[83] or caveman[84] lawyer persists, even if this type of anachronism is, in fact, an ethical violation waiting to happen.[85] But even attorneys who "only touch a computer under duress,

[81] *See id.*

[82] *See id.*

[83] *See* Debra Cassens Weiss, *Lawyers Have Duty to Stay Current on Technology's Risks and Benefits, New Model Ethics Comment Says*, ABA Journal Law News (Aug. 6, 2012, 7:46 PM) http://www.abajournal.com/news/article/lawyers_have_duty_to_stay_current_on_technologys_risks_and_benefits/, *archived at* https://perma.cc/WPZ4-2DYH.

[84] *See Unfrozen Caveman Lawyer*, SATURDAY NIGHT LIVE TRANSCRIPTS, http://snltranscripts.jt.org/91/91gcaveman.phtml, *archived at* https://perma.cc/M7GB-DGJZ ("Sometimes when I get a message on my fax machine, I wonder: 'Did little demons get inside and type it?' I don't know! My primitive mind can't grasp these concepts.") (last visited Apr. 5, 2016).

[85] *See* Megan Zavieh, *Luddite Lawyers Are Ethical Violations Waiting to Happen*, LAWYERIST.COM (last updated July 10, 2015), https://lawyerist.com/71071/luddite-lawyers-ethical-violations-waiting-happen/, *archived at* https://perma.cc/6V4W-94J7.

and take comfort in paper files and legal research from actual books"[86] must deal with technology.[87] The adequate practice—or perhaps simply "the practice" of law does not exist without technology, and there is no longer a place for lawyers who simply "hope to get to retirement before they need to fully incorporate technology into their lives."[88]

[41] "Really?" goes the refrain. "Why can't I just practice the way I always have, without [insert mangled, vaguely-recognizable technology portmanteau] getting in the way?"

[42] Well, for one thing, to the extent attorneys rely on the protections of privilege to serve their clients, said attorneys must understand how the confidentiality of their communications and work product may be compromised by the technology they use. Technologies introduce complexity that, in turn, may affect privilege—especially when "many lawyers don't understand electronic information or have failed to take necessary precautions to protect it."[89] But how much understanding,

[86] Lois D. Mermelstein, *Ethics Update: Lawyers Must Keep Up with Technology Too, American Bar Association – Business Law Today*, BUSINESS LAW TODAY (Mar. 2013), http://www.americanbar.org/publications/blt/2013/03/keeping_current.html, *archived at* https://perma.cc/T8CF-ZWND.

[87] *See* Blair Janis, *How Technology Is Changing the Practice Of Law*, GP SOLO, http://www.americanbar.org/publications/gp_solo/2014/may_june/how_technology_changing_practice_law.html, *archived at* https://perma.cc/23P5-PGM7 (last visited Apr. 5, 2016).

[88] Kevin O'Keefe, *We Need Laws Requiring Lawyers to Stay Abreast of Technology?* LEXBLOG: ETHICS & BLOGGING LAW (Mar. 28, 2015), http://kevin.lexblog.com/2015/03/28/we-need-laws-requiring-lawyers-to-stay-abreast-of-technology/, *archived at* https://perma.cc/8DR5-XK43.

[89] *Attorney-client Privilege: Technological Changes Bring Changing Responsibilities for Attorneys and Legal Departments,* CORPORATE LAW ADVISORY, http://www.lexisnexis.com/communities/corporatecounselnewsletter/b/newsletter/archive/2014/01/06/attorney-client-privilege-technological-changes-bring-changing-responsibilities-for-attorneys-and-legal-departments.aspx, *archived at* https://perma.cc/XQ53-P3MF (last visited Apr. 5, 2016).

exactly, may be required to competently represent clients in matters concerning E-Discovery, or data security, or even privacy? At many organizations, "[p]rivacy issues get handled by anyone who wants to do them" because the subject matter area is understaffed or ignored.[90] The key technological issues relevant to E-Discovery versus data privacy may be somewhat different, but the "solutions" companies find are eerily similar: the practitioners that are actually doing the work are often those who have been delegated the work, whose "expertise" is somewhat home-grown and may, in fact, not really represent true technological competence at all.[91]

[43] What, then, are the requirements for expertise? Perhaps a pragmatic approach is best. Certainly, practitioners who use technology—again, likely all of them—must take some well-defined, initial steps toward acquiring the appropriate skill set. This might be as straightforward as the lawyer familiarizing herself with the relevant technologies at issue. Although it may sound a bit *too* easy, "just being well-versed enough to understand the issues is a big plus."[92] That being said, "those considering a career in cybersecurity or privacy will need to spend time developing some level of technical expertise."[93] In short, the answer is "it depends"

---

[90] Daniel Solove, *Starting a Privacy Law Career*, LINKEDIN PULSE (Aug. 27, 2013), https://www.linkedin.com/pulse/20130827061558-2259773-starting-a-privacy-law-career?forceNoSplash=true, *archived at* https://perma.cc/G78L-DM2X.

[91] *See* Peter Geraghty & Sue Michmerhuizen, *Think Twice Before You Call Yourself an Expert*, YOUR ABA (Mar. 2013), http://www.americanbar.org/newsletter/publications/youraba/201303article11.html, *archived at* https://perma.cc/HJK7-RSLG .

[92] Solove, *supra* note 90.

[93] Alysa Pfeiffer-Austin, *Four Practical Tips to Succeed in the Cybersecurity and Privacy Law Market*, ABA Security Law (Dec. 9, 2015), http://abaforlawstudents.com/2015/12/09/four-practical-tips-to-succeed-in-the-cybersecurity-and-privacy-law-market/, *archived at* https://perma.cc/AH9A-JCTU.

and "no one really knows – yet." In this relatively new space, actual decisions and definitive standards for "technological competence" are thin on the ground. Below we will examine some of the relevant rules and guidelines to consider.

A. Recent Guidelines in the Ethics Rules

[44] Most attorneys do not have specialized training focused on a particular technological field. Certainly the vast majority do not hold themselves out as experts in cybersecurity, cloud-based storage, social media, biometrics, or any of a variety of related disciplines. However, even in the absence of expertise, there are some basic ethical rules that provide a framework for determining a practitioner's professional duties and obligations with regard to technology—specifically, rules pertaining to competent client representation, adequate supervision, confidentiality, and communications.[94]

1. Competent Client Representation (Model Rule 1.1)

[45] As discussed briefly above, almost four years ago, the America Bar Association formally approved a change to the Model Rules of Professional Conduct to establish a clear understanding that lawyers have a duty to be competent not only in the law and its practice, but also with respect to technology. Detailed below, the passage of this rule contemplated changes in technology and eschewed specifics. Rather than a paint-by-numbers approach, ABA Model Rule 1.1 puts the responsibility on attorneys to understand their own—and their clients'—needs, and how new technologies impact their particular practice.

[46] ABA Model Rule 1.1 states that:

---

[94] *See* David G. Ries, *Cybersecurity for Attorneys: Understanding the Ethical Obligations*, LAW PRACTICE TODAY (Mar. 2012), http://www.americanbar.org/publications/law_practice_today_home/law_practice_today_archive/march12/cyber-security-for-attorneys-understanding-the-ethical-obligations.html, *archived at* https://perma.cc/N4VM-N4NG.

> A lawyer shall provide competent representation to a client. Competent representation requires legal knowledge, skill, thoroughness and preparation reasonably necessary for the representation.[95]

[47] ABA Model Rule 1.1 was amended in 2012 by Codified Comment 8 as follows:

> To maintain the requisite knowledge and skills, a lawyer should keep abreast of changes in the law and its practice, *including the benefits and risks associated with relevant technology*, engage in continuing study and education and comply with all continuing legal education requirements to which the lawyer is subject.[96]

[48] Some note that Rule 1.1 "does not actually impose any new obligations on lawyers;"[97] neither does it require perfection.[98] Instead it "simply reiterates the obvious, particularly for seasoned eDiscovery lawyers, that in order for lawyers to adequately practice, they need to understand the means by which they zealously advocate for their clients."[99] One article noted, in fact, that Comment 8 was evidence of "the ABA's desire to nudge lawyers into the 21st century when it comes to

[95] MODEL RULES OF PROF'L CONDUCT R. 1.1 (2014).

[96] MODEL RULES OF PROF'L CONDUCT R. 1.1 cmt. 8 (2014) (emphasis added).

[97] Jenson, Watson & Sherer, *supra* note 40, at 2.

[98] *See* James Podgers, *You Don't Need Perfect Tech Knowhow for Ethics' Sake—But a Reasonable Grasp Is Essential*, ABA JOURNAL (Aug. 9, 2014), http://www.abajournal.com/news/article/you_dont_need_perfect_tech_knowhow_for_ethics_sake--but_a_reasonable_grasp, *archived at* https://perma.cc/CB3P-R7YL.

[99] Jenson, Watson & Sherer, *supra* note 40, at 2.

technology."[100] It did, however, caution that it was "a very gentle nudge."[101]

[49] Nudge or not, that message has resonated across the United States. In the four years since that amendment was approved and adopted by the ABA, twenty-one states since have adopted the ethical duty of technological competence for lawyers.[102] As for many of the states that have not formally adopted the change to their Model Rules of Professional Conduct, those may still explicitly or implicitly acknowledge this emerging duty to be competent in technology, having a basic understanding of technologies their clients use, and a duty to keep abreast of such changes including a required awareness of regulatory requirements and privacy laws.[103]

---

[100] Kelly H. Twigger, Symposium, *Ethics in Technology and eDiscovery – Stuff You Know, but Aren't Thinking About*, ARK. L. REV. (Oct. 16, 2014), http://law.uark.edu/documents/2014/10/TWIGGER-Ethics-in-Technology-and-eDiscovery.pdf, *archived at* https://perma.cc/LTG8-7AYU.

[101] *Id.*

[102] These states are: Arizona, Arkansas, Connecticut, Delaware, Idaho, Illinois, Iowa, Kansas, Massachusetts, Minnesota, Nebraska, New Hampshire, New Mexico, New York, North Carolina, Ohio, Pennsylvania, Utah, Virginia, West Virginia, and Wyoming. *See* Robert Ambrogi, *20 States Have Adopted Ethical Duty of Technological Competence*, LAW SITES (Mar. 16, 2015), http://www.lawsitesblog.com/2015/03/11-states-have-adopted-ethical-duty-of-technology-competence.html, *archived at* https://perma.cc/B5TF-D6NJ (last updated Dec. 23, 2015) (listing 20 states not including Nebraska); *see also Basic Technology Competence for Lawyers*, Event Details, NEBRASKA BAR ASSOC. (Apr. 6, 2016), https://nebar.site-ym.com/events/EventDetails.aspx?id=788239&group=, *archived at* https://perma.cc/SMU6-58TU ("[T]he need to be aware of and have a working knowledge of technology…is ethically required of all lawyers.").

[103] Ann M. Murphy, *Is It Safe? The Need for State Ethical Rules to Keep Pace with Technological Advances*, 81 FORDHAM L. REV. 1651, 1659, 1665–66 (2013), http://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=4876&context=flr, *archived at* https://perma.cc/V69A-EETR.

2. Supervision (Model Rules 5.1 and 5.3)

[50] ABA Model Rule 5.1 also bears on a lawyer's duties regarding technology insofar as duties aided or supported by technology are performed by someone other than the attorney. This responsibility extends to immediate as well as remote support staff, with ABA Model Rule 5.1 requiring that "[l]awyers must also supervise the work of others to ensure it is completed in a competent manner."[104] This attempt at establishing "the principle of supervisory responsibility without introducing a vicarious liability concept"[105] has led to considerations regarding inexperience generally,[106] but the implications for technological applications should be clear—an associate or other paralegal professional is much more likely to use technology to support legal work[107] than she is to make a representation before a court or like body.

[51] ABA Model Rule 5.3 also sets forth responsibilities of partners and supervising attorneys to non-lawyer assistants. This set of ethical considerations further reinforces the responsibilities attorneys have to apply sufficient care in their practice when outsourcing supporting legal

---

[104] Samantha V. Ettari & Noah Hertz-Bunzl, *Ethical E-Discovery: What Every Lawyer Needs to Know*, LEGALTECHNEWS (Nov. 10, 2015), http://www.kramerlevin.com/files/Publication/d7dec721-693a-4810-a4b9-32dfe9c1864b/Presentation/PublicationAttachment/018a444a-d7de-46b2-bc16-506cff88d346/EDiscovery-Legaltech%20News11.10.15..pdf, *archived at* https://perma.cc/4YMR-XL9U (referring to MODEL RULE OF PROF'L CONDUCT 5.1).

[105] AMERICAN BAR ASSOCIATION, A LEGISLATIVE HISTORY: THE DEVELOPMENT OF THE ABA MODEL RULES OF PROFESSIONAL CONDUCT, 1982-2005 560 (2006).

[106] Jeffrey P. Reilly, *Rule 5.1 of the Rules of Professional Conduct: What Must Corporate General Counsel Do?* ASSOCIATION OF CORPORATE COUNSEL, BALTIMORE CHAPTER FOCUS 2Q12 5–6 (2012), http://www.milesstockbridge.com/pdf/publications/ReillyACCArticle.pdf, *archived at* https://perma.cc/G26J-NTJE.

[107] *See* Jennifer Ellis, *What Technology Does a Modern US Lawyer Generally Use in Practice?*, QUORA (Mar. 22, 2014), https://www.quora.com/What-technology-does-a-modern-US-lawyer-generally-use-in-practice, *archived at* https://perma.cc/4FX4-2UV7.

work to inexperienced non-professionals, and to ensure that confidentiality is maintained with outsourcing staff.[108] This is not just a matter of supervising specific tasks. It also contemplates knowing which tasks are appropriate for delegation, both within the firm and to third-party vendors. For example, if a delegate of the attorney uses technology to begin an engagement, it's possible that such an arrangement could be viewed as "establish[ing] the attorney-client relationship," which may be prohibited under ABA Model Rule 5.5.[109]

3. Duty of Confidentiality (Model Rule 1.6)

[52] ABA Model Rule 1.6 states that it is critical that lawyers do not reveal confidential or privileged client information.[110] When information was kept in an attorney's head, or perhaps committed to a sheet of paper, historical precedent on how to comply with this duty may have been helpful. In the "world of tomorrow,"[111] looking to the past for answers makes little sense, especially in those instances where the attorney is unclear as to how information is stored, accessed, maintained, or utilized.

[53] Model Rule 1.6 also considers a duty of confidentiality that resides at the core of every attorney's role and serves as one of the attorney's most important ethical responsibilities. Model Rule 1.6 generally defines the duty of confidentiality as follows: "A lawyer shall not reveal information relating to the representation of a client unless the client gives informed consent, the disclosure is impliedly authorized in order to carry out the

[108] *See* MODEL RULES OF PROF'L CONDUCT R. 5.3.

[109] Frances P. Kao, *No, a Paralegal Is Not a Lawyer*, ABA BUS. LAW TODAY, (Jan./Feb. 2007), https://apps.americanbar.org/buslaw/blt/2007-01-02/kao.shtml, *archived at* https://perma.cc/3J2N-ELPA.

[110] *See* MODEL RULES OF PROF'L CONDUCT R. 1.6.

[111] *See* Jon Snyder, *1939's 'World of Tomorrow' Shaped Our Today*, WIRED (Apr. 29, 2010, 8:00 PM), http://www.wired.com/2010/04/gallery-1939-worlds-fair/, *archived at* https://perma.cc/D5V4-36R5.

representation or the disclosure is permitted [elsewhere]."[112]

[54] This rule is broad. It encompasses any client information, confidential or privileged, shared or accessible to the attorney and is not limited to just confidential communications. Further, it may only be relinquished under the most onerous of circumstances.[113] A lawyer shall not, therefore, reveal information relating to the representation of a client unless the client gives informed consent, the disclosure is impliedly authorized in order to carry out the representation, or the disclosure is permitted elsewhere in the rules.

[55] In 2000, the Advisory Committee looked into its crystal ball and considered ESI on various platforms, in different repositories, in various forms. It then added Comment 18 to Rule 1.6, requiring reasonable precautions to safeguard and preserve confidential information. Comment 18 states that, "[A] lawyer [must] act competently to safeguard information relating to the representation of a client against … inadvertent or unauthorized disclosure by the lawyer or other persons who are participating in the representation of the client or who are subject to the lawyer's supervision."[114] Indeed, "[p]artners and supervising attorneys are required to take reasonable actions to ensure that those under their supervision comply with these requirements."[115]

---

[112] MODEL RULES OF PROF'L CONDUCT R. 1.6.

[113] *See* Saul Jay Singer, *Speaking of Ethics: When Tarasoff Meets Rule 1.6*, WASHINGTON LAWYER (May 2011), https://www.dcbar.org/bar-resources/publications/washington-lawyer/articles/may-2011-speaking-of-ethics.cfm, *archived at* https://perma.cc/A7E4-DSH6.

[114] MODEL RULES OF PROF'L CONDUCT R. 1.6 cmt. 18.

[115] David G. Ries, *Cybersecurity for Attorneys: Understanding the Ethical Obligations*, LAW PRACTICE TODAY (Mar. 2012), http://www.americanbar.org/publications/law_practice_today_home/law_practice_today_archive/march12/cyber-security-for-attorneys-understanding-the-ethical-obligations.html, *archived at* https://perma.cc/59Q2-55Q4.

[56] In addition to the ABA's commentary, state and local professional organizations have issued guidance as well. In establishing a specific roadmap for lawyers to attain the skills necessary to meet their ethical obligations with respect to relevant technology in the practice of law, and returning to the California Bar's Formal Opinion 2015-193, there is a sort of checklist that may assist lawyers in meeting their ethical obligations to develop and maintain core E-Discovery competence in the following areas:[116]

- Initially assessing E-Discovery needs and issues, if any;
- Implementing or causing (the client) to implement appropriate ESI preservation procedures, ("such as circulating litigation holds or suspending auto-delete programs");[117]
- Analyzing and understanding the client's ESI systems and storage;
- Advising the client on available options for collection and preservation of ESI;
- Identifying custodians of potentially relevant ESI;
- Engaging in competent and meaningful meet and confers with opposing counsel concerning an E-Discovery plan;
- Performing data searches;
- Collecting responsive ESI in a manner that preserves the integrity of the ESI; and
- Producing responsive, non-privileged ESI in a recognized and appropriate manner.

[57] But this technological competence inherent in the Duty of Competence represents only one third of the ethical duties that govern an

---

[116] *See* State Bar of Cal. Standing Comm. on Prof'l Responsibility and Conduct, Formal Op. 2015-193, 3–4 (2015) [hereinafter Cal. Ethics Op. 2015-193] (discussing what an attorney's ethical duties are in the handling of discovery of electronically stored information).

[117] Ettari & Hertz-Bunzl, *supra* note 104.

attorney's interaction with technology. This ESI and litigation skills checklist does *not* address "the scope of an attorney's duty of competence relating to obtaining an opposing party's ESI;"[118] nor does it consider the skills required of non-litigation attorneys, which must be inferred from the rule.

[58] In addition, the State Bar of California's Standing Committee on Professional Responsibility and Conduct, Formal Opinion 2010-179 states that "[a]n attorney's duties of confidentiality and competence require the attorney to take appropriate steps to ensure that his or her use of technology in conjunction with a client's representations does not subject confidential client information to an undue risk of unauthorized disclosure."[119]

[59] In reference to the duty of confidentiality, the New York County Lawyer's Association's Committee on Professional Ethics examined shared computer services amongst practitioners in Opinion 733, noting that an "attorney must diligently preserve the client's confidences, whether reduced to digital format, paper, or otherwise. The same considerations would also apply to electronic mail and websites to the extent they would be used as vehicles for communications with the attorney's clients."[120] The New York State Bar's Committee on Professional Ethics Opinion 842 further stated that, when "a lawyer is on notice that the [client's] information…is of 'an extraordinarily sensitive nature that it is reasonable to use only a means of communication that is completely under the

[118] Cal. Ethics Op. 2015-193, *supra* note 116, at fn. 7.

[119] State Bar of Cal. Standing Comm. on Prof'l Responsibility and Conduct, Formal Op. 2010-179, 7 (2010) (discussing whether an attorney violates the duties of confidentiality and competence she owes to a client by using technology to transmit or store confidential client information when the technology may be susceptible to unauthorized access by third parties).

[120] N.Y. Cnty. Lawyers' Ass'n Comm. on Prof'l Ethics, Formal Op. 733, 7 (2004) (discussing non-exclusive referrals and sharing of office space, computers, telephone lines, office expenses, and advertising with non-legal professionals).

lawyer's control,…the lawyer must select a more secure means of communication than unencrypted Internet e-mail.'"[121]

4. Communications (Model Rule 1.4)

[60] ABA Model Rule 1.4 on Communications also applies to the attorney's use of technology and requires appropriate communications with clients "about the means by which the client's objectives are to be accomplished," including the use of technology.[122]

[61] In construing all of these Model Rules and comments, it is clear that attorneys who are not tech-must (1) understand their limitations; (2) obtain appropriate assistance; (3) be aware of the areas in which technology knowledge is essential; and (4) evolve to competently handle those challenges; or (5) retain the requisite expert assistance. This list applies equally to data security issues, such as being aware of the risks associated with cloud storage, cybersecurity threats, and other sources of potential harm to client data, and can easily be extended to include awareness and understanding with respect to domestic and foreign data privacy issues.

[62] The ethical obligations to safeguard information require reasonable security, not absolute security. Accordingly, under such rules and related guidance from the Proposal from the ABA Commission on Ethics 20/20,[123] the factors to be considered in determining the reasonableness of

---

[121] N.Y. State Bar Ass'n Comm. on Prof'l Ethics, Formal Op. 842 (2010) (discussing using an outside online storage provider to store client's confidential information).

[122] MODEL RULES OF PROF'L CONDUCT R. 1.4 (1983); *see also* 204 PA. CODE § 81.4 (1988), http://www.pacode.com/secure/data/204/chapter81/chap81toc.html, *archived at* https://perma.cc/6FG5-9VP3 (incorporating ABA Model Rule 1.4 into Pennsylvania's Model Rule 1.4).

[123] *See* ABA Comm. on Ethics 20/20, *Introduction and Overview* (Feb. 2013), http://www.americanbar.org/content/dam/aba/administrative/ethics_2020/20121112_ethics_20_20_overarching_report_final_with_disclaimer.authcheckdam.pdf, *archived at* https://perma.cc/D2ZY-NYEU.

the lawyers' efforts with respect to security include:

> (1) The sensitivity of the information;
> (2) The likelihood of disclosure if additional safeguards are not employed;
> (3) The cost of employing additional safeguards;
> (4) The difficulty of implementing the safeguards; and
> (5) The extent to which the safeguards adversely affect the lawyer's ability to represent the client.[124]

As New Jersey Ethics Opinion 701 states, "[r]easonable care however does not mean that the lawyer absolutely and strictly guarantees that the information will be utterly invulnerable against all unauthorized access. Such a guarantee is impossible."[125]

B. Ethics and Social Media

[63] When considering their ethical duties with respect to technology, lawyers today must confront a host of challenges that would have been almost unimaginable even ten years ago. The rise and proliferation of social media as a daily part of most people's personal and professional lives has created one such challenge.[126] Numerous courts have

---

[124] MODEL RULES OF PROF'L CONDUCT R. 1.6(c) cmt. 18 (1983).

[125] Opinion 701 also highlights, if inadvertently, the challenges attorneys face when trying to modify existing practices to fit new technologies. As part of the inquiry underpinning Opinion 701's guidance, the opinion notes that "nothing in the RPCs prevents a lawyer from archiving a client's file through use of an electronic medium such as PDF files or similar formats." This note is nearly laughable when read in the context of current practice, as it suggests that attorneys were (or are?) concerned about whether PDF files are appropriate for retaining paper documents. N.J. Advisory Comm. on Prof'l Ethics, Formal Op. 701 (2006), https://www.judiciary.state.nj.us/notices/ethics/ACPE_Opinion701_ElectronicStorage_12022005.pdf, *archived at* https://perma.cc/EV9H-BN3T.

[126] *See* Brian M. Karpf, *Florida's Take on Telling Clients to Scrub Social Media Pages*, LAW 360 (Sept. 15, 2015, 4:33 PM), http://www.law360.com/articles/702288/florida-s-

addressed—and continue to address—attorney duties with respect to social media in the context of spoliation motions when social media evidence has been lost, destroyed, or obfuscated due to negligence, or in accordance with attorney advice.[127] In addition, given the novelty and complexity of the issues, and in the interest of consistency, state bar associations have begun to address issues associated with attorney use of, counseling on, and preservation of social media.

[64] The Association of the Bar of the City of New York's Committee on Professional and Judicial Ethics, in Formal Opinion 2010-2, provided some helpful guidelines on attorney access to social media, stating that "[a] lawyer may not use deception to access information from a social networking webpage," either directly or through an agent.[128] While focused on behaviors that attorneys and their agents should not undertake when developing a case, the opinion does note that the "potential availability of helpful evidence on these internet-based sources makes them an attractive new weapon in a lawyer's arsenal of formal and informal discovery devices," and also offers up "the Court of Appeals' oft-cited policy in favor of informal discovery."[129] Simply put, the duty is twofold: an attorney must both be aware of social media and know how to use social media to provide effective representation.

2. State Bar Association Guidance

[65] State bar associations are becoming increasingly involved in

take-on-telling-clients-to-scrub-social-media-pages, *archived at* https://perma.cc/NZ3W-FHPS.

[127] *See id.*

[128] N.Y.C. Bar Ass'n Comm. on Prof'l. Ethics, Formal Op. 2010-2 (2010), http://www.nycbar.org/ethics/ethics-opinions-local/2010-opinions/786-obtaining-evidence-from-social-networking-websites, *archived at* https://perma.cc/JT9K-2EGV (discussing lawyers' obtainment of information from social networking websites).

[129] *Id.*

providing guidance on social media and its implications for the practice of law. For example, in 2014, the New York and Pennsylvania State Bar Associations and the Florida Professional Ethics Committee issued guidance on social media usage by attorneys and addressed the obligations of attorneys to understand how various platforms work, what information will be available to whom, the ethical implications of advising clients to alter or change social media accounts, and the value of ensuring adequate preservation of social media evidence.

i. New York

[66] The Social Media Ethics Guidelines of the Commercial and Federal Litigation Section of the New York State Bar Association provide specific guidance for the use of social media by attorneys.[130] Guideline 4, relating to the review and use of evidence from social media, is divided into four subparts, all of which provide specific and pertinent guidance to attorneys:

- Guideline No. 4.A: Viewing a Public Portion of a Social Media Website, provides that "[a] lawyer may view the public portion of a person's social media profile or public posts even if such person is represented by another lawyer. However, the lawyer must be aware that certain social media networks may send an automatic message to the person whose account is being viewed which identifies the person viewing the account as well as other information about such person."[131]

- Guideline No. 4.B: Contacting an Unrepresented Party to View a Restricted Portion of a Social Media

[130] Mark A. Berman, Ignatius A. Grande & James M. Wicks, *Social Media Ethics Guidelines of the Commercial and Federal Litigation Section of the New York State Bar Association*, THE NEW YORK STATE BAR ASSOCIATION (June 9, 2015), http://www.nysba.org/socialmediaguidelines/, *archived at* https://perma.cc/4ZSN-BXT4.

[131] *Id.*

> Website, provides that "[a] lawyer may request permission to view the restricted portion of an unrepresented person's social media website or profile. However, the lawyer must use her full name and an accurate profile, and she may not create a different or false profile to mask her identity. If the person asks for additional information from the lawyer in response to the request that seeks permission to view her social media profile, the lawyer must accurately provide the information requested by the person or withdraw her request."[132]

- Guideline No. 4.C: Viewing A Represented Party's Restricted Social Media Website, provides that "[a] lawyer shall not contact a represented person to seek to review the restricted portion of the person's social media profile unless an express authorization has been furnished by such person."[133]

- Guideline No. 4.D: Lawyer's Use of Agents to Contact a Represented Party, "as it relates to viewing a person's social media account," provides that "[a] lawyer shall not order or direct an agent to engage in specific conduct, or with knowledge of the specific conduct by such person, ratify it, where such conduct if engaged in by the lawyer would violate any ethics rules."[134]

ii. Florida

---

[132] *Id.*

[133] *Id.*

[134] *Id.*

[67] In Advisory Opinion 14-1, the Florida Bar Association's Professional Ethics Committee confirmed that an attorney could advise a client to increase privacy settings (as so to conceal from public eye) and remove information relevant to the foreseeable proceedings from social media as long as an appropriate record was maintained—the data preserved—and no rules or substantive laws regarding preservation and/or spoliation of evidence were broken.[135]

iii. Pennsylvania

[68] In 2014, the Pennsylvania Bar Association issued a Formal Opinion that included detailed guidance regarding an attorney's ethical obligations with respect to the use of social media. Among other guidelines, the Opinion specifically stated that:

- Attorneys may advise clients about the content of their Social networking websites, including the removal or addition of information;
- Attorneys may connect with clients and former clients;
- Attorneys may not contact a represented person through social networking websites;
- Although attorneys may contact an unrepresented person through social networking websites, they may not use a pretextual basis for viewing otherwise private information on social networking websites; and
- Attorneys may use information on social networking websites in a dispute.[136]

[135] *See* Fla. State Bar Comm. on Prof'l Ethics, Proposed Op. 14-1 (2015), http://www.floridabar.org/TFB/TFBResources.nsf/Attachments/B806500C941083C785257E730071222B/$FILE/14-01%20PAO.pdf?OpenElement, *archived at* https://perma.cc/DK9W-A44Z.

[136] Pa. Bar Ass'n. Comm. on Ethics, Formal Op. 2014-300, 2 (2014), http://www.americanbar.org/content/dam/aba/events/professional_responsibility/2015/M

3. ABA Model Rule 3.4

[69] Finally, although ABA Model Rule 3.4 on Fairness to Opposing Party and Counsel does not directly address social media, the principles behind the rule apply in the social media context. The Rule provides that an attorney shall not "unlawfully obstruct another party's access to evidence or unlawfully alter, destroy or conceal a document or other material having potential evidentiary value" nor shall the attorney "counsel or assist another person" to undertake such actions.[137]

C. Guidance on Duties Related to Cybersecurity

[70] As we discussed above in Section II, attorneys face a complex threat landscape when it comes to security concerns related to the protection of their clients' data.[138] Although the scope of an attorney's ethical obligations in this regard remains somewhat unclear, there are several sources of guidance relevant to how lawyers are expected to manage cybersecurity risks.

[71] One such source that squarely addresses the issue is the Resolution issued by the ABA's Cybersecurity Legal Task Force. The Resolution contains a detailed Report explaining the ABA's position regarding the growing problem of intrusions into computer networks utilized by lawyers and law firms, and urges lawyers and law firms to review and comply with the provisions relating to the safeguarding of confidential client information.[139] As the ABA noted in its Report, defending the

ay/Conference/Materials/pa_formal_op_2014_300.authcheckdam.pdf, *archived at* https://perma.cc/G6EY-PBFF.

[137] MODEL RULES OF PROF'L CONDUCT R. 3.4 (1983).

[138] *See supra* Part II.

[139] *See* ABA Cybersecurity Legal Task Force, Resolution 118, 2 (August 2013), http://www.americanbar.org/content/dam/aba/administrative/law_national_security/resolution_118.authcheckdam.pdf, *archived at* https://perma.cc/UQ44-3Q2C.

confidentiality of the lawyer-client relationship and preservation of privilege in communications and attorney work product are fundamental to public confidence in the legal system.[140] Attorneys are directed to (1) keep clients reasonably informed as set forth in the Model Rules of Professional Conduct, as amended in August 2012 and adopted in the jurisdictions applicable to their practice; and (2) comply with other applicable state, federal, and court rules pertaining to data privacy and cybersecurity.[141] The ABA further urges the respect and preservation of the attorney client relationship during the pendency of any actions in which a government entity aims to deter, prevent, or punish unauthorized, illegal intrusions into computer systems and networks used by lawyers and law firms.

[72] The comment to ABA Model Rule 5.7 states, perhaps somewhat axiomatically, that when "[a] lawyer performs law-related services or controls an organization that does so, there exists the potential for ethical problems."[142] This, combined with Model Rule 1.6's requirement for attorneys to safeguard and protect client information, suggests further potential duties associated with cybersecurity.[143] As one author notes

> Fulfillment of a law firm's duty to maintain client confidences in today's world of cyberattacks requires much more than legal knowledge and legal skills. It requires sophisticated computer knowledge and skills far beyond legal practice. That is why cybersecurity experts should be used to assist in any law firm's client's data protection

---

[140] *See id.* at 4.

[141] *See id.* at 16.

[142] MODEL RULES OF PROF'L CONDUCT R. 5.7, cmt. 1 (1983).

[143] *See* MODEL RULES OF PROF'L CONDUCT R. 1.6.

> efforts.[144]

Indeed, "[t]raining in security, including cybersecurity should be a part of every lawyer's education. It is especially important for lawyers who do electronic discovery".[145]

[73] On a related subject, in Formal Opinion 2015-3, the New York City Bar Association issued guidance indicating that lawyers do *not* violate their ethical duties by reporting suspected cybercrime to law enforcement.[146] If an attorney has performed "reasonable diligence" to determine whether a prospective client is actually attempting fraud, the opinion says, then the attorney is free to report.[147] The Opinion continued, highlighting the lack of duty associated with individuals who are not actually clients, stating that an

> attorney who discovers that is he the target of an Internet-based trust account scam does *not* have a duty of confidentiality to the individual attempting to defraud him, and is free to report the individual to law enforcement authorities, because that person does not qualify as a prospective or actual client of the attorney.[148]

---

[144] Ralph C. Losey, *The Importance of Cybersecurity in eDiscovery*, E-DISCOVERY LAW TODAY (May 9, 2014) http://www.ediscoverylawtoday.com/2014/05/the-importance-of-data-security-in-ediscovery/, *archived at* https://perma.cc/P64J-NYQ7.

[145] Ralph C. Losey, *The Importance of Cybersecurity to the Legal Profession and Outsourcing as a Best Practice – Part Two*, E-DISCOVERY TEAM (May 18, 2014), http://e-discoveryteam.com/2014/05/18/the-importance-of-cybersecurity-to-the-legal-profession-and-outsourcing-as-a-best-practice-part-two/, *archived at* https://perma.cc/W3HW-AHCC.

[146] N.Y.C. Bar Ass'n Comm. on Prof'l Ethics, Formal Op. 2015-3, 4–5 (2015), http://www2.nycbar.org/pdf/report/uploads/20072898-FormalOpinion2015-3-LAWYERSWHOFALLVICTIMTOINTERNETSCAMS.pdf, *archived at* https://perma.cc/6BHV-V2YC.

[147] *Id.* at 1.

[148] *Id.* at 6 (emphasis added).

V. CONCLUSION

[74] It goes without saying that we live (and work) in interesting times. Cloud technology offers convenience, flexibility, cost savings—and a host of potential security issues that existing "hard-copy world" rules aren't fit to address. The details of top-secret corporate transactions are now hashed out on collaborative virtual platforms that may be vulnerable to damage, destruction, or unauthorized access. And the increasing ubiquity of social media makes it ever more likely that lawyers and clients alike may post information without appreciating the potential legal ramifications. New technologies have the capacity to enrich our personal lives and enhance our professional lives, but they also create complex and novel challenges for lawyers already subject to a web of ethical duties concerning competence and confidentiality.

[75] Given the speed with which this dynamic area is changing, the issues raised in this piece may well feel dated within months of publication as the next new product or service revolutionizes another fundamental aspect of human interaction and connectivity. Nevertheless, in this article we have outlined some of the many challenges facing attorneys operating in a threat-laden high-tech landscape, taken a look at the ways in which existing and emerging ethical rules and guidelines may apply to the practice of law in the digital age, and opened a door to further conversation about all of these issues as they continue to evolve.

AI INTEGRATION WITH BLOCKCHAIN

***Please analyze the following readings separately in one paragraph only at the end of your three-page assignment**

a. Bernard Marr, "Artificial Intelligence And Blockchain: 3 Major Benefits of Combining These Two Megatrends" https://www.forbes.com/sites/bernardmarr/2018/03/02/artificial-intelligence-and-blockchain-3-major-benefits-of-combining-these-two-mega-trends/#3b322d954b44
b. Francesco Corea, "The Convergence of AI and Blockchain: What's the Deal?" (this one is slightly more complex) https://medium.com/@Francesco_AI/the-convergence-of-ai-and-blockchain-whats-the-deal-60c618e3accc