A Study into the Layers of Automated Decision Making:

Emergent Normative and Legal Aspects of Deep Learning

Argyro P Karanasiou

Centre for Intellectual Property Policy & Management, Faculty of Media & Communication, Bournemouth University, Bournemouth, United Kingdom

Dimitris A Pinotsis

The Picower Institute for Learning and Memory and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Boston MA, United States

Corresponding author: akaranasiou@bournemouth.ac.uk

Argyro P Karanasiou is a Senior Lecturer in Law at Bournemouth University & a Visiting Research Fellow at the ISP Centre at Yale Law School.

Dimitris A Pinotsis is a Visiting Scientist at the Massachusetts Institute of Technology & an Honorary Senior Research Associate at University College London.

A Study into the Layers of Automated Decision Making:

Emergent Normative and Legal Aspects of Deep Learning

The paper dissects the intricacies of Automated Decision Making (ADM) and urges for refining the current legal definition of AI when pinpointing the role of algorithms in the advent of ubiquitous computing, data analytics and deep learning. ADM relies upon a plethora of algorithmic approaches and has already found a wide range of applications in marketing automation, social networks, computational neuroscience, robotics, and other fields. Whilst coming up with a toolkit to measure algorithmic determination in automated/semi-automated tasks might be proven to be a tedious task for the legislator, our main aim here is to explain how a thorough understanding of the layers of ADM could be a first good step towards this direction: AI operates on a formula based on several degrees of automation employed in the interaction between the programmer, the user, and the algorithm; this can take various shapes and thus yield different answers to key issues regarding agency. The paper offers a fresh look at the concept of "Machine Intelligence", which exposes certain vulnerabilities in its current legal interpretation. To highlight this argument, analysis proceeds in two parts: Part 1 strives to provide a taxonomy of the various levels of automation that reflects distinct degrees of Human – Machine interaction and can thus serve as a point of reference for outlining distinct rights and obligations of the programmer and the consumer: driverless cars are used as a case study to explore the several layers of human and machine interaction. These different degrees of automation reflect various levels of complexities in the underlying algorithms, and pose very interesting questions in terms of regulating the algorithms that undertake dynamic driving tasks. Part 2 further discusses the intricate nature of the underlying algorithms and artificial neural networks (ANN) that implement them and considers how one can interpret and utilize observed patterns in acquired data. Finally, the paper explores the scope for user empowerment and data transparency and discusses attendant legal challenges posed by these recent technological developments.

Keywords: machine learning algorithms; ANN; automation; personhood; algorithmic accountability.

> "I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted."
>
> — Alan Turing, Computing machinery and intelligence (1950)

1. Going Underwater: On Submarines Swimming in Different Strokes

In 1973 the Science Research Council (SRC) commissioned Sir James Lighthill, a Professor of Applied Mathematics at Cambridge, to write a report assessing the progress of AI research in the UK. The Lighthill report (SRC 1973) outlined three separate yet intertwined areas within the broad scope of AI research:

(i) Advanced Automation (namely, specific automated tasks performed by machines such as pattern recognition),

(ii) Computer Based Research (namely, computational simulations modelling neurophysiological theories) and,

(iii) Robotics (namely, automatic devices that mimic human functions).

Lighthill's findings, painted a somewhat pessimistic picture of the potential that the new –at the time- field of robotics might have to operate autonomously. Due to the complexity of the tasks such fully automated systems would have to face, human involvement would always be required. Simpler programs "written to perform in highly specialised problem domains, when the programming takes very full account of the results of human experience and human intelligence" might perform well in specific tasks; building an autonomous robot on the other hand, relies on "general-purpose programs seeking to mimic the problem-solving aspects of human CNS activity over a rather wide field."(SRC 1973). Even so, the human element still cannot be fully taken out of the equation. This,

Lighthill posits, is due to the fact that the highly sophisticated datasets used in automated systems present the system with a "combinatorial explosion", namely a wealth of possible states of a system. This can only be dealt with resorting to heuristics, "whereby it is the programmer's intelligence that helps the machine deal with the combinatorial explosion". As a result, it would be unrealistic to expect a "highly generalised system that can handle a large knowledge base effectively in a learning or self-organising mode" to be developed in the near future.

Lighthill's ominous predictions have proven to be untrue. Since 1973 great advances have occurred in machine learning research, which has led to a wide range of application in everyday life: Virtual Personal Assistants like Apple's Siri or Microsoft's Cortana, driverless cars and smart thermostats are only a few examples to a rapidly expanding list. An important component of these applications is Automated Decision Making (ADM), that is, the ability of algorithms to provide solutions in tasks with ambiguous outcomes and determine the optimal among a set of possible answers. In light of these developments, this paper attempts to provide an overview of the various layers of algorithmic determinism in automated and semi-automated tasks. Our hope is that this analysis could serve as a useful point of reference for further techno-legal research in autonomous systems.

Fast forward to 2016, Microsoft released an artificial application into the online social sphere: a ChatBot called Tay.ai, which was designed to interact with Twitter users and learn from these interactions. Within 24 hours, Microsoft had to deactivate Tay's Twitter account, due to a large amount of retweets of racism comments on Tay's feed, often including further offensive commentary by the ChatBot (Perez 2016). Although such racial commentary is not unusual online (Williams et al, 2016), the case of Tay is of particular interest given that it provides empirical evidence of advanced forms of AI that

is able to mimic human behaviour. This interaction between the machine and the human is an intricate process that includes various degrees of automation. These result from mixing together the user feedback with the algorithm's behaviour.

The question of how *real* and *simulated* intelligence measure up in AI is hardly a new one (for a good overview see Haugeland 1985). Note for example Chomsky's reading of the Alan Turing test (Turing 1950) as an approach that separates the cognitive from the biological elements in order to provide an answer as to whether machines can be perceived by humans as able to think, not different to fooling someone into believing the "submarines can swim" (Chomsky, 1996). This, Chomsky concludes, is a "question of decision, not a question of fact", not different to fooling someone into believing the "submarines can swim".

This interpretation of "*intelligence*" lies at the heart of the argument put forth here: to legally assess Automated Decision Making, one needs to go beyond the realm of biological and cognitive abilities and consider the essence of the concept of "personhood": what defines a person and when is a person autonomous? In other words, the level of autonomy displayed by the agent or the machine will also determine the level of liability, which is currently a puzzling notion for legal scholars addressing AI. To highlight this point, the paper uses driverless cars as a case study and explains how fully automated systems bestow upon us the task to develop our theorizing in order to accommodate artificial agents within legal doctrines. As it will be shown in the remainder of the paper, the matter of "intelligence" in AI is not merely of philosophical nature but its definition is much needed to provide solid grounding for emergent legal issues, such as tortious liability (Chopra & White, 2011). The latter is of course a legal convention, which provides us with a safe tool to address challenging issues in automated systems (i.e. liability in driverless cars) but is not on its own enough to account for the

reconfiguration of key concepts, such as causation and responsibility. This explains the focal point of this paper, which revolves around the personhood of artificial agents.

Moving away from Chomsky's narrow interpretation of the Turing test, Russell and Norvig (2003) draw an interesting distinction between an artefact's behaviour and an artefacts pedigree: "we can conclude that in some cases, the behaviour of an artefact is important, while in others it is the artefact's pedigree that matters. Which one is important in which case seems to be a matter of convention. But for artificial minds, there is no convention."

To elucidate such intricacies, the following section provides an overview of ADM and its mechanics, namely some related machine learning algorithms and the current trend towards deep learning.

2. A Contextual Analysis of Emergent Normative and Legal aspects in Automated Systems: The Intricacies of Machine Learning Algorithms

The aim of this section is to first establish an understanding of the technical context, within which ADM occurs. This will not only allow us to explain better how a definition of "intelligence" in AI is somewhat elusive but it will also provide a solid methodological grounding, given that the approach taken here is a techno-legal overview of automated systems. Recent advances in machine learning and computational complexity theory have been further boosted by the ability to collect, manipulate and store vast amounts of data. ADM is a natural product of these exciting developments and has found a wide range of applications in seemingly unrelated fields like marketing automation, social networks, computational neuroscience, robotics, banking, transportation and others.

Machine learning algorithms often employ artificial neural networks (ANNs). This means

that the computational units these algorithms use to perform intelligent functions resemble biological networks and neurons. ANNs take advantage of powerful algorithms that are trained using large datasets available in many industries (image databases, security or healthcare records, traffic or consumer behaviour data, online platform analytics, etc.) so that they can correctly decide upon suitable actions when new data are presented to them in a similar way to what a human agent would do; for example to recognise faces or operate driverless cars. The purpose of ADM is to be able to act without the need of human intervention. They are be able to deal with novel conditions, that is take the right decision even when the dataset presented to them is different from the one they have been trained on, e.g. a driverless car should be able to navigate in a road it has not had access before.

How do ANN algorithms learn to perform complicated tasks efficiently? Put simply, the answer lies in exploiting both increased computational power and vast amounts of data already collected. This data is used by the programmer to train the algorithm. Technically, training is often done in one of the following three ways: supervised, unsupervised or reinforcement learning, see e.g. (Mohri et al., 2012). These are technical terms that relate to the details of the training process and are distinct from potential interactions with the user after the algorithm is passed on to her in e.g. human-in-the-loop and similar applications.

Supervised learning (SL) occurs when during training the algorithm is fed with both an input and the correct decision (output). For example, when the algorithm has to distinguish between faces and objects in a scene, the input would be an image and the output a class index, e.g. 1 for faces and 2 for objects. The algorithm is then given pairs of images and class indices that are used to fine tune its parameters. The algorithm has to find the correct class index when – after learning- it is presented with a new image that

may or may not contain a face (Nakajima et al, 2000).

Unsupervised learning (UL) is quite similar conceptually. Using the above simple example, the difference is that the algorithm would have to guess whether the image contains a face or not without being explicitly given the corresponding indices during the training process (Kumar et al., 2010). Of course, when designed, the algorithm is fed with some information about the task, e.g. it would know it should decide between two possible alternatives, however it is not given which images contain faces and which do not, it has to discover these differences based on certain features that the images might contain, e.g. eyes, nose and mouth at close proximity in all images that contain faces. In a more difficult scenario, the algorithm might even have to decide how many classes or categories there might be in the data, something that might lead to it over- or under-estimating this number. In such clustering or classification tasks the algorithm puts together points that are related in some conceptual space. Of course, the dimensions of this space (which features should be selected) are crucial for making the algorithm efficient and are chosen by the programmer in the design stage. This is important as it might introduce a bias in the output of the decision process: depending on what features the programmer chooses to be important, the algorithm might take different decisions. We call this the "bias" introduced by the programmer to the ADM algorithm. The reader should keep this term in mind as we will come back to it in section 4.2 below. Bias is not only an issue in unsupervised learning but also in other machine learning approaches like Reinforcement Learning to which we now turn:

Reinforcement learning (RL) is slightly more complicated: it decouples actions from rewards and the algorithm aims not at taking the "right" action (decision), but maximizing the reward it receives (Sutton and Barto, 1998). This is merely a technical distinction that renders the description of the relevant algorithms slightly more complicated – for

example, the algorithm might have to take several actions one after the other to maximize an end goal (reward). Interestingly, this decoupling speaks to the ability of the algorithm to take sequential decisions that are related to each other and think ahead in time; for example, the DeepMind algorithm that plays the Atari game Breakout should find a balance between the time it spends at each location firing and the speed it moves if it wants to accumulate sufficient reward (high score) and successfully proceed to the next level (Mnih et al., 2015). Furthermore, this balance might change in time or as the level of the game advances. Contrary to the other two approaches, the emphasis in RL is in combining several decisions (or actions) to get the most benefit out of them. In other words, reward is a complicated function of two or more decisions that might be unknown even to the programmer, let alone the user herself.

RL is today considered to be a promising avenue for building intelligent algorithms that can adapt to different environments and even tasks; an important limitation in older machine learning approaches was the lack of flexibility: e.g. an algorithm might learn to play chess at master level but would be unable to play checkers, which for most human players that know the rules chess would be easy to pick up. This is why algorithms are often trained to perform within a limited set of conditions and cannot succeed when rules changes, even slightly. In a paper published last year, DeepMind researchers showed that the same algorithm could perform well in several Atari games without being trained in each one individually (Minh et al., 2015) Essentially, the algorithm learns different mappings between actions and rewards online and is able to flexibly maximize the benefit it receives when the environment (game) changes.

All three learning approaches have a long history in machine learning, however recent successes like the DeepMind algorithm for playing Atari games discussed above followed technical advances sometimes referred to collectively as Deep Learning (DL). For

example, the DeepMind work uses Deep-Q Learning which is a combination of RL and DL (Van Hasselt et al., 2015). Roughly speaking, the term "Deep" here refers to increasing the power (and complexity) of an algorithm by taking its basic constituent parts and using them recursively, that is feeding the output of one part to the other. Crucially, each part uses a similar learning process, however only after combining all parts together is the system (building a deep architecture) able to perform well. If the architecture of the algorithm is changed, e.g. a smaller number of constituent parts are used, then the algorithm might not be able to take the right decision of find the action that maximise its reward.

Architectural details like e.g. the exact number of parts (layers) in the system or how "big" each part should be in terms of how many computational units should be used are often found by experience. This is in contrast with older approaches and rule-based simulations where the algorithms were implemented in much smaller computer infrastructures and the role of different computational elements involved was more transparent. Interestingly, it might not be a principled explanation as to why certain deep (extended) architectures work and others don't something often referred to as the deep algorithms being somehow "opaque". This idea has its roots in neuroscience where a succession of brain areas – e.g. the ventral system- plays a similar role to a deep network architecture. In this setting, certain brain areas situated away from sensory regions light up and respond to different stimuli e.g. some areas respond to faces and others to objects. This means that these areas are sensitive to the category of the visual stimuli and can distinguish between categories. Crucially, earlier (visual) areas would respond to anything placed in the visual field regardless of its category. However, only higher areas that receive input from several upstream regions are able to distinguish between different categories of visual stimuli. In brief, the brain decides about the category of the stimulus

by combining signals from several areas that interact in a large network. Similarly, it is only after the programmer endows its algorithm with several parts and builds a "deep" hierarchical architecture that the algorithm can distinguish between classes of visual stimuli.

So what have we lost by making the algorithm deep? Maybe we have found a way to replace humans with intelligent agents that can perform well and take the right decisions; however, we cannot claim that the algorithm really understands or interprets its input the way a human would do. This poses an interesting challenge for law, and in particular regarding the concept of "agency", as deep algorithms have the ability to *act* upon their input, e. g. take a decision. In this case, the definition of "act" is *stretched* beyond the narrow confinements of conventional legal formalism; algorithms do not serve as mere tools but are able to take well informed decisions under little or no supervision at all.

Most importantly, there exists an additional dimension that further muddles the waters for legally assessing ADM: what is the scope for the user's involvement in the decision process? Given the complexity in the process of decision making, a clear understanding of the interactions between the machine and the human agent is necessary not only for attributing responsibility for the outcome of the decision met but further to explore the causality, intent and risk assessment. Take for example the law of negligence, a tort introduced partly in response to the problems of agency: direct liability would only apply in supervised systems, whereas indirect liability under the doctrine of *respondeat superior* would require a certain level of foreseeability, namely "normalised expectations for the technical capacities of computer action" (Teubner, 2007).

In applications that require a human-in-the-loop like Brain Computer Interface (BCI), assisted Decision Making and Health Informatics the user already plays an active role in this process. In such cases, the user acts supplementary to the algorithm and interacts with

it. This leads to increased performance and efficiency of the algorithm and good performance even in situations of high uncertainty or increased risk. What makes human-in-the-loop algorithms different to autonomous systems is not the way training is carried out but the possibility of human intervention at intermediate stages of the training process. The human intervenes to enhance the algorithm's performance by bringing in knowledge the algorithm has no access to. Intermediate training follows the general procedures we have described above but the user has a decisive role in selecting new training datasets that have been pre-processed by her, e.g. throw irrelevant parts away or intervene at intermediate stages to assess the quality of results produced and guide the algorithm accordingly. For example, in (Awasthi et al.,2015) an algorithm used limited supervision to cluster data in a certain number of groups with the help of the user who at each stage told the algorithm whether it should split or merge some of it .

Thus far we have discussed the technical details underlying machine learning algorithms used in ADM. These summarise what we earlier called the artefact's pedigree. In the following section, we focus on the artefact's behaviour and use driverless cars as a case study to explore the various levels of automation: this allows us to gain a better understanding of various degrees of human-machine interaction, which will serve as a reference point for the remainder of the paper and shall aid us in our quest to understand the balance between the algorithm's inner workings – that are often opaque – and human intervention.

3. A Taxonomy of Automation Layers: Driverless Cars as a Case Study

The prospect of fully autonomous vehicles "designed to be capable of safely competing journeys without the need for a driver" (Department for Transport Code of Practise) has

certainly gained momentum in the past few years: Google Chauffeur software currently tested in autonomous vehicles in California, Rio Tinto's autonomous haulage systems operating since 2008 in Australia or Volvo's pioneering programme "Drive me" expected to release autonomous vehicles to customers in Gothenburg by 2017 are a few indicative cases of the great potential automated systems have shown in the transport industry (Atkins 2015). This however is far from removing drivers completely "off the loop", although many manufacturers have already introduced semi-automated vehicles with driving assistance features, such as controlling the brake, throttle and steering, supporting active lane-keeping or using sensors to deliver full speed adaptive cruise control (KPMG 2013).

It is thus apparent that automated systems, such as autonomous vehicles, operate on several different degrees of automation, according to how much control is yielded to the driver. In other words, the novel element here is not automation *per se* but the variety of degrees of interaction between the man and the machine. Take for example the case study of driverless cars explored here: automated driving is not really a striking fact nowadays; the auto-mobile started replacing the horse-drawn carriages in the turn of the 20th century. The initial scepticism towards the new risks posed by the technological advances was followed by gradual adoption of the new means of transport, mainly due to the codification of automated driving in law. As Moris (2007) notes "In the 1890s improvements in the internal combustion engine, legal and political developments which severely restricted the power of cities to regulate the types of traffic on their streets (won by bicycle advocates), the [aforementioned] invention of traffic rules, and smooth new asphalt street surfaces paved the way for the private automobile. Enticed by high speeds, point-to-point travel and the flexibility to roam across the urban landscape, the public adopted the new innovation in droves". Transport related legal issues, mainly liability,

have been dealt with a dynamic body of regulations at a national and international level, which have taken an anthropocentric view: assumption of risk, bad judgement, and reasonable foreseeability, are a few grounds upon which causality can be established. At the same time, they all have one common point of departure: human error as a sine qua non of the decision making process.

The elimination of human error is however also one of the key elements behind self-driving cars. A 2008 NHTSA report attributes 40% of collisions to "recognition errors", caused by distractions, and 35% to "decision errors", such as speeding. It is thus expected that removing the human element from driving will enhance road safety (NHTSA, 2008). Recent progress in computer vision like the use of massively parallel graphic processing units and deep learning algorithms have led to a revolution in the field of driverless cars. The quest for self-driving vehicles was initiated with DARPA's Grand Challenges: this was a competition among such vehicles where external operators were allowed to intervene in the vehicles' route to minimize risk and ensure safety (e.g. by stopping and restarting the vehicles). Since then, several milestones have been reached and fully autonomous driving has become a reality (Urmson et al., 2008; Levinson et al., 2011; 2014; Wei et al., 2013). Of course, due to the complexity and breadth of possible driving conditions, achieving fully autonomous cars that have sufficient training so that they are able to perform well in any situation is far from solved (despite using huge training datasets, that include millions of highway and road images etc.). However, extending basic computer vision algorithms to the level of replacing human agents is now considered viable and several reports of self-driving cars have appeared in the media, e.g. (Rosen, R.,2012; Hull,L., 2013).

Thus, it is not the technology or the externalities it unavoidably creates that hinder our legal understanding of automated decision making. What is challenging for legal minds,

is an unprecedented variety of interfaces and levels of interaction between the human and a machine learning algorithm. To put it differently, to fully assess these algorithms one will have to perceive to what extent the human element (directly by human-in-the-loop interventions or indirectly at the design stage) is present in the "intelligence" demonstrated by the algorithm. As noted in section 2 above, it is imperative that a basic taxonomy for ADM is adopted prior to any legal evaluation to enhance our understanding of how each "automated" task involves constant shifts of roles from executing to merely supervising (Sheridan 1970).

The study of these interactions has given rise to many theories discussing ontological and deontological approaches regarding automated functions and the degree of human involvement (Fitts 1951). As a result, many taxonomies of various degrees of automation have been suggested in a quest to localise informational control in the human or automaton domain: Sheridan and Verplank's ten degrees of automation (1978) are probably the most widely adopted theory that describes variations of control from human to collaborative and to fully automated, Endsley and Kaber's theory (1999) emphasizes on supported, blended or automated decision making, whereas Riley's taxonomy (1989) uses a mixed assessment based on various levels of autonomy that intersect with different degrees of intelligence. These theories have provided the ground for authorities such as the NHTSA or the Society of Automobile Engineers (SAE, see figure 1) to identify 5 levels of automation in computer assisted driving:

(i) No-Automation (Level 0), i.e. the system automatically assists the driver to regain lost control of the vehicle.

(ii) Function-specific Automation (Level 1), i.e. the system controls one function.

(iii) Combined Function Automation (Level 2), i.e. the system controls at least two functions.

(iv) Limited Self-Driving Automation (Level 3), i.e. the driver cedes full control under specific conditions,

(v) Full Self-Driving Automation (Level 4), i.e. the driver is not expected to become involved throughout the duration of the trip.

Level	Name	Narrative definition	Execution of steering and acceleration/ deceleration	Monitoring of driving environment	Fallback performance of *dynamic driving task*	System capability (*driving modes*)	BASt level	NHTSA level
***Human driver* monitors the driving environment**								
0	**No Automation**	the full-time performance by the *human driver* of all aspects of the *dynamic driving task*, even when enhanced by warning or intervention systems	Human driver	Human driver	Human driver	n/a	Driver only	0
1	**Driver Assistance**	the *driving mode*-specific execution by a driver assistance system of either steering or acceleration/deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	Human driver and system	Human driver	Human driver	Some driving modes	Assisted	1
2	**Partial Automation**	the *driving mode*-specific execution by one or more driver assistance systems of both steering and acceleration/deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	**System**	Human driver	Human driver	Some driving modes	Partially automated	2
***Automated driving system* ("system") monitors the driving environment**								
3	**Conditional Automation**	the *driving mode*-specific performance by an *automated driving system* of all aspects of the *dynamic driving task* with the expectation that the *human driver* will respond appropriately to a *request to intervene*	System	**System**	Human driver	Some driving modes	Highly automated	3
4	**High Automation**	the *driving mode*-specific performance by an *automated driving system* of all aspects of the *dynamic driving task*, even if a *human driver* does not respond appropriately to a *request to intervene*	System	System	**System**	Some driving modes	Fully automated	3/4
5	**Full Automation**	the full-time performance by an *automated driving system* of all aspects of the *dynamic driving task* under all roadway and environmental conditions that can be managed by a *human driver*	System	System	System	**All driving modes**	-	

Source: cyberlaw.stanford.edu/loda – SAE Information Report J3016

Further to this, the NHTSA Federal Automated Vehicles Policy published in September 2016 by the US Department of Transportation, outlines in more detail the term "highly automated vehicle" (HAV), which represents SAE Levels 3-5 vehicles with automated systems that are responsible for monitoring the driving environment. This variety of human – machine interaction introduces a new complexity: "the vehicle must be capable of accurately conveying information to the human driver regarding intentions and vehicle performance", as well as to its environment, namely "other external actors with whom the HAV may have interactions (other vehicles, pedestrians, etc.)". To put

this differently, it does matter whether the average observer can tell whether a vehicle is autonomous or not, as this changes the degree of reliance towards the ability of a driver to manoeuvre and shapes reasonable expectations accordingly. This is particularly interesting when one considers Level 3 SAE systems, which are expected to be monitored by the driver, although human capacity to stay alert when disengaged from the driving task may be limited.

Driverless cars are a recent example where automated systems have made great progress and reached a level, where the operator can be completely ignored. Earlier examples include aviation (Spizer, 1987) and medicine (Thompson, 1994), leading up to the emergence of the DoNotPay Bot in 2016, the world's first "robot lawyer", offering free legal advice to the homeless. We have chosen to discuss driverless cars in the paper, as the various degrees of automation discussed above, capture perfectly this interplay between the operator and the agent. As Sheridan notes "Automation has moved from open-loop mechanization of industrial revolution, then to simple closed loop linear control, then to non-linear and adaptive control and recently to a mic of crisp and fuzzy rule-based decision, neural nets and generic algorithms that truly recognize patterns and learn" (Sheridan 2000). This in turn has also marked a shift from automated ML (aML) to interactive ML (iML) (Holzinger, 2016), namely an almost seamless interaction between the machine and the operator. The more sophisticated the system is, the more it changes the nature of human performance, challenging thereby our understanding of who the operator of a given task is, and to what extent she needs to apply own cognitive capacities (Parasuraman, 1997). From a legal standpoint, this is highly problematic as such interactions lend anthropomorphic traits to otherwise automatically executed tasks. In a similar vein, Calo (2015) outlines three distinctive features in robotics that blend the boundaries between the human and the machine: embodiment of the algorithm (e.g. the

car in our case study), emergence (the "coupling of complexity and usefulness") and social valence, namely the public reliance on automated systems. Ultimately, he concludes that new juridical insights will be required to fully perceive this emerging field from a legal viewpoint and accurately evaluate to what extend automated systems can be treated as social actors, able to "*think*" for us after having benefited from our social experiences. This echoes Teubner (2007), who having reviewed Luhmann and Latour, explains how most legal actors are created by social attribution, without the need to possess any ontological human properties, such as reflexive capacities or empathy. That said, artificial agents are still beyond the narrow confinements of our current anthropocentric view of legal actors.

Can autonomous cars drive us, in the same sense that submarines can swim? So far we have focused on how advances in machine learning have led to highly sophisticated automated systems that can potentially throw the operator out-of-the-loop. To understand this better, let us take the Google driverless car as an example and focus on how it can operate with minimal supervision. The Google algorithm for driverless cars performs the following operations: (i) self-localization using 3D map technologies (ii) determination of static and moving obstacles (iii) classification of information/objects by using machine vision (iv) generation of road condition predictions (v) evaluation of these predictions against real circumstances (vi) automated actions like steering, braking or accelerating, if required (Titiriga, 2016). These are the same operations a human driver would have to undertake; however the sense of agency is in this case different: what do notions like "average reasonable person", "free will", "mens rea" and degrees of culpability mean in the case of driverless cars? Such questions present us with an "indirect agency", a status which is not easy to assess legally using frequently evoked criteria.

Let us then consider each of the above steps independently: in operations (ii) and (iii) the algorithm has to perform image and object recognition, segmentation and classification. Given the limited degree of automation in the decision making process, it can be argued that these steps correspond to levels 0-2, in the SAE taxonomy mentioned above. In other words, the algorithm has to first understand how many objects exist in its view and then classify them into pedestrians, cars, traffic lights etc. This means that the algorithm has to boost interesting parts of the image over not so interesting ones; for example, be able to distinguish between a pedestrian standing next to a still or obscure background, e.g. a traffic light at a crossing or in a pavement with low lighting. Segmentation is then carried out using some sensors (cameras, lasers etc.) that should be able to learn new environments in an unsupervised way (Levinson, J., & Thrun, S., 2014). In this context, recognition and classification of human and objects in the car's proximity might go beyond simple processing of visual input through the car's camera and applying labels to objects using a database stored in the car: they might require autonomous interactions with electronic systems and databases outside the vehicle like GPS-based guidance systems and information from the Department of Transportation (DOT) that would allow the algorithm to localize the vehicle and its neighbouring objects and surroundings (Zhu,J., et al., 2014). Furthermore, information about the car's location and other parameters (speed, direction etc.) should be passed on to a central (global) guidance system and database at a remote location, e.g. DOT so that other (neighbouring) vehicles might be informed about the car's trajectory and parameters.

Operations (iv)-(vi) above are more complicated and as such, correspond to SAE levels 3 – 5 (see Figure 1 above): on top of image processing and computer vision tasks, the algorithm of the driverless car has to solve an inherently dynamic problem where on top of image processing the algorithm has also to predict trajectories in time, both its own

and neighbouring cars e.g. predict the future location of the car in the front given its speed to avoid collision in case it breaks unexpectedly. It also has to generate appropriate steering commands, breaking, acceleration and be able to associate past and future driving conditions, e.g. if the ground map includes information about a congested road coming up the algorithm could look for alternative routes or try to slow down even though obstacles might not be directly visible. All these operations endow the algorithm with a novel sense of agency as it effectively acts in lieu of a driver and behaves like one. What are the criteria for legally assessing this new sort of agency?

This question does not suggest that automated vehicles operate on a legal vacuum. On the contrary, the issue of liability has been debated many times at a national, federal and international level and although incoherent, most solutions suggested in the regulatory domain move towards strict liability. Given however the different types of driverless cars (reflecting various shades of automation), there is no size that fits all: Volvo, for instance has declared that the company will pay for any damages caused by its fully autonomous IntelliSafe Autopilot system. With regards to Google's car, the National Highway Traffic Safety Administration (NHTSA) has recognized that the software, not the human, is the driver. At the same time though, the international Vienna Convention on Road Traffic gives responsibility for the car to the driver, requiring that "[e]very driver shall at all times be able to control his vehicle". The amendment to Vienna Convention, which came into effect in 2016, to include article 8 paragraph 5bis VC, does little in clarifying matters regarding autonomous vehicles: as it is premise on the assumption that such automated systems can be overridden by the driver, it does not take into account fully automated systems. Far from establishing legal certainty, the current regulative framework regarding automated vehicles is still dispersed and in working progress. At the same time, the issue of agency is barely addressed, mainly due to the challenging issue

of proving actual causation in automated technology (Wittenberg, 2016). Next, follows an attempt to understand the agent's artificial "intelligence" through the lens of personhood – a doctrinal approach beyond the strict confines of liability.

4. Deep Learning Conundrums: The Importance of Assumptions in Automated Decision Making Algorithms.

To address this question we will examine below the concept of personhood together with algorithmic transparency. But before, let us pause for an intermediate summary: so far, we have attempted to provide a descriptive (section 2) and normative analysis (section 3) of machine learning algorithms. These analyses have validated the hypothesis set out in the introduction, that ADM is a challenging concept for law because it rests on both the artefact's pedigree (see section 2) and the artefact's behaviour (see section 3). These are two separate yet intertwined elements in the process of mimicking human behaviour. In the case of driverless cars considered above, it was shown how human behaviour reinforces the artefact's pedigree, while at the same time the artefact's behaviour can occur without any human involvement. Therein lies the heart of the argument put forth here: the understanding of what robotic "intelligence" is by legal scholars is often limited; to this shortcoming one should add the increased complexity of modern techniques like RL and deep algorithms in AI that lead to a difficult conundrum; importantly, this conundrum cannot be addressed purely with metaphors as it is often the case for other questions that are new to legal research (Calo 2016). Earlier, we considered different levels of automation in machine learning algorithms and different shades of human agency inbuilt in systems using deep learning. This led us to conclude that tools for legal assessment that are currently available (e.g. Vienna Convention) are expected to be

unable to capture the different levels of automation and human-machine interaction. For example, RL is often characterised by an opaque mechanism of decision making: although RL robots bear anthropomorphic features, it is still not clear to the lawmaker how to deal with this emergent concept of "assimilated personhood"[1]. In this final part, the paper explores the necessity for a new concept of personhood together with algorithmic transparency in ADM and attempts to show how modern machine learning algorithms like RL present us with new challenges that require novel sets of standards.

4.1.Artificial Personhood v. Simulated Personhood: Focusing on "the loop"

(Gray 1921) defined personhood as the quality of as any entity possessing "intelligence and will". The idea that AI systems should be given entitlements to personhood is hardly a new one: there is already rich literature (Allan and Widdison, 1996; Kerr and Millar, 2001; Chopra and White, 2011) that suggests that autonomous artificial agents could potentially be considered as entities meriting "legal" personhood.

This is not the first time that entities other than a person are entitled to the responsibilities and rights associated with the notion of personhood. In the early 19th century, the US Supreme Court in *Dartmouth* described corporations as "an artificial being, invisible, intangible, and existing only in contemplation of the law", which displays in fact certain personhood virtues, not as a person but as a "mere creature of law." (Dartmouth College v. Woodward, 17 U.S. 518, 636 (1819). Since then, modern corporate law has developed a more nuanced approach, acknowledging that these entities

[1] The term is used here to highlight how this is different not only to the traditional "personhood" but also to the notion of "artificial personhood" (doctrine of corporate personhood).

- being the creation of private initiative and market forces- incorporate competing interests that need to be accounted for (Kaeb 2015). In a similar vein, robots and artificial agents are highly automated systems that are equally premised on "private initiative and market forces" and would therefore fit the criteria of "legal personhood" as such. In the era of algorithms being the driving force behind unmanned systems that could inflict harm, like military drones, it is imperative not to afford them "the blessings of perpetual life and limited liability" (Rehnquist dissenting in *Pellotti* with regard to banking corporations).

This proposition has of course not gone without criticism: automated systems cannot experience life as a good to itself given their lack of consciousness (Aleksander 1994; Franklin 1995) and would fall beyond the strict confinements of liability as a punishment aiming at deterrence (Bentham, 2009). Such arguments however oversimplify the way in which automated systems operate and do not carefully consider the various levels of automation, as described above. Solum (1992) has therefore disregarded these claims as purely "behaviouristic approaches" and has urged for a distinction between *simulated* and *artificial* intelligence. This would be a good first step towards addressing some of the most complicated regulatory problems posed by AI: limited foreseeability of actions, operations based on a highly compartmentalised and opaque design, and a narrow scope of controlled tasks, are only a few examples that demonstrate the need to fully grasp the contours of "intelligence" in AI (Scherer, 2016).

4.2. The "Intelligence and Will" in Deep Learning: An interpretation of opacity

We saw earlier, that deep learning algorithms for ADM have an intricate architecture, are often opaque and allow for various levels of human-machine interaction and autonomy.

In other words, they are much more complex and less transparent than earlier rule- based algorithms, however, this additional complexity has not adequately been taken into account in their legal assessment to date. We also suggested that such intricacies render the understanding the concept of "personhood" associated with ADM algorithms problematic.

Earlier, we associated personhood with any entity possessing "intelligence and will". A highly sophisticated and automated system can be considered to possess "personhood" but in what ways is the system "intelligent" and has "will"? Furthermore, the system was designed by a programmer and might sometimes be influenced by the user. Both the programmer and the user have their one distinct "personhoods", so how do they interfere with the "system's personhood"?

We here propose that to address the above difficult questions one needs to adopt a legal approach that will focus on both what the infrastructure and behaviour of the automated system is *and* what the role of the human element (programmer, user) might be, see also (Jones, 2015). This means that one needs to go beyond older approaches that put too much emphasis on how (i) efficient (*cf* Citron, 2007) and (ii) objective the algorithm is (Zarsky, 2015) without at the same time considering what the potential role of the human influence might be. As we saw earlier, this influence can be important for the algorithms output; for example, it might introduce *biases* in the outputs of the automated decision process.

Dissecting the role of the human element is not an easy task, because, as we saw earlier, human influence might be hidden behind opaque architectures of the sort used in deep learning or might be indirect in the case of human-in-the-loop applications. This might be important for the correct legal assessment of liability and similar issues in modern ADM: if one neglects the influence of the programmer or operator, she runs the chance

of not correctly attributing to humans flaws in the ADM algorithms for which the humans should be held responsible. Of course, the opacity of the algorithms does not render this an easy task especially for legal scholars; however only by taking a deeper look into the ADM mechanics could we have any hope of properly understanding concepts like personhood and liability associated with highly automated systems.

A good number of scholars (Pasquale, 2015; Citron and Pasquale, 2014; Crawford and Schultz, 2014; Zarsky, 2016) are currently focusing their critique towards the high levels of opacity and urge the law to "open the black box of algorithms" or even set up a body of independent auditors to carefully examine ADM (Sandvig et al., 2014). In section 2 above, we saw that one important aspect of this opacity that can perhaps be easily quantified is the "bias" introduced by the programmer to the ADM algorithm: this referred to some feature selection or similar process that crucially affects the output (decision) of the algorithm and which results from the programmer's direct input at the stage of designing the algorithm. We agree with the aforementioned scholars about the need to restore transparency as a much needed ex post measure to eliminate bias and evaluate human involvement and liability. Yet, we will argue, opening the black box of algorithms only sees part of the picture when it comes to modern ADM algorithms as it merely focus on the algorithms' *design*. On the other hand, the "intelligence and the will" of the algorithm cannot be disconnected from its performance after the design process (and training) has been finalised: for example, when the driverless car has to navigate in real world conditions and interact with human agents (imagine such a car navigating through a street filled with other cars driven by humans). At that moment, the algorithm has its own personhood, mimics human behaviour and perhaps continuously interacts with humans like a normal person would do. All these are emergent normative features that should be taken into careful consideration during proper legal assessment of deep

learning algorithms: we argue that understanding the mechanics of these algorithms at the stage (level) of their design is insufficient and should be supplemented by the study of what the overall scope of human involvement at all stages might be including training and unsupervised or semi –supervised performance. For example, consider a driverless car that is first trained in a racing track, then performs successfully in the highway and then is assisted by a human when navigating in narrower streets. Is it enough to merely study the technical details of the algorithms that are used and also try to embed morality in their design? We argue it is not, and suggest that the law should also attempt to define the "intelligence" or "smartness" (Hildenbrandt, 2015) of the algorithm as well as how this is affected by the subsequent human influence (after the algorithm is designed and training has been completed).

5. Conclusion: From the Imitation Game to the *Voigt-Kampff* Test - Towards an Updated Legal Understanding of Machine Intelligence

This paper has attempted to provide a normative and legal grounding of the "intelligence" demonstrated in automated systems that rely on deep learning. This is highly relevant nowadays, as the technological advances in robotics and cognitive sciences have paved the way to more sophisticated systems that can act and in a completely autonomous manner. These systems demonstrate remarkable abilities to mimic human behaviour: this can be happen in such unprecedented ways that interactions between algorithms and humans can be quite difficult to predict, e.g. consider Microsoft 2016's apology on their official blog regarding their Chabot Tay, and its racist comments on Twitter. The law has therefore to inevitably adopt a new concept of personhood that will deal with behaviours

of modern human-like agents. This concept should go beyond the scope of traditional (weak) AI and reconsider what "personhood" might be; also, how personhood can be described when human-like autonomous agents that act in an "intelligent" manner, learn and evolve on their own interact with humans in real world environments.

This unavoidably takes us down the treacherous road of providing definitions of concepts like "intelligence"; a tedious task in itself due to the relativity the concept bears. A simple question that comes to mind when one first tries to define this concept is the following: is it a concept that can be understood in terms of a *mechanism* (or an algorithm) that generates certain (human-like) behaviours or is it a matter of a human *perceiving* an agent (a human or a machine) as intelligent? Although Turing's original intention in 'Computing Machinery and Intelligence' was to explore whether a computer can "imitate a brain" (Copeland, 2004), he then admitted to be sceptical as to how the intelligence of a machine was to be perceived: "The extent to which we regard something as behaving in an intelligent manner" he noted (Turing, 1950) "is determined as much by our own state of mind and training as by the properties of the object under consideration" (see also Minsky, 1988 for a similar view). In other words, Turing suggests that "intelligence" relates to how we perceive it in a manner remarkably similar to how the legal system operates: Turing's "perception" of intelligence is akin to the principle of "interpretation". The legal system tries to interpret human behaviours *not* to understand the mechanisms (algorithms) that might have generated them; this might be one reason why automated systems are not easily perceived in law and humanities in general. To address these shortcomings, theorists have sought to elucidate additional dimensions of machine intelligence, like consciousness (Floridi, 2005), along the same lines of the empathy test employed in Philip Dick's fictitious Voigt-Kampff test (Dick, 1968). Whereas intelligent processing shall always be opaque, it is desirable to go past the prima facie

anthropomorphism of automated systems and actually enhance our understanding of what their "intelligence" might be. Deep Learning for instance, might yield results that even the programmers cannot anticipate. We therefore suggest that our perception of machine intelligence should be enhanced; this could either happen *ex ante* ("at the input stage") or *ex post* ("at the output stage"):

(i) *ex ante* efforts could include monitoring or prescribing the algorithm's design features and principles e.g. carefully selecting training data or initial weights so that they are consistent with legal or ethical constraints.

(ii) *ex post* efforts on the other hand, refer mostly to the user's interpretation and feedback after the algorithm has performed an intelligent function (taken a decision). This is also important as it places ADM within the socio-legal context it belongs to.

Machine learning has reached such a sophisticated level that it could not only result in misrepresenting an automated system that passed the Turing test as a human but importantly escape liability due to the judiciary's inability to attribute a concept of "personhood" to the system (algorithm). In his response to the Lighthill report we mentioned at the beginning, Prof Longuet Higgins made a remarkably timely remark, relevant to our discussion thirty years later. Interestingly, Prof Higgins foresaw the danger of an algorithmic system wrongfully escaping liability: "The mathematician's ability to discover a theorem, the formulation of a strategy in master chess, the interpretation of a visual field as a landscape with three cows and a cottage, the feat of hearing what someone says at a cocktail party and the triumph of reading one's aunt's handwriting, all seem to involve the same general skill, namely the ability to integrate in a flash a wide range of knowledge and experience. Perhaps Advanced Automation will

indeed go its own sweet way, regardless of Cognitive Science; but if it does so, I fear that the resulting spin-off is more than likely to inflict multiple injuries on human society".

Since the Lighthill report, the concept of machine "intelligence" has no doubt gained new dimensions. The law however seems to be lagging behind. This paper has sought to explore the challenges put forth by the application of modern machine learning algorithms like deep networks and reinforcement learning in the area of Automated Decision Making (ADM), which merits further research and consideration. We hope that our findings shall mobilise legal scholars and ethicists to undertake the difficult task of further dissecting the emergent normative features associated with ADM in the not so distant future.

AKNOWLEDGEMENTS

The authors wish to thank Joseph Savirimuthu for all his support and hard work put into this special issue as well as Roger Brownsword, the ISP Centre Fellows at Yale Law School, the Hariri Institute at Boston University (especially Azer Bestavros and Ran Canetti), and the anonymous reviewers for their valuable feedback. The usual disclaimer applies.

BIBLIOGRAPHY

Aleksander I. 1994. "Towards a Neural Model of Consciousness." *Proceedings ICANN 94*. Berlin: Springer

Allan, T., and R. Widdison. 1996. "Can computers make contracts?" *Harvard Journal of Law and Technology*, no 9: 25–52.

Atkins R. 2015. *"Connected and Autonomous Vehicles: Introducing the future of mobility."* White Paper.

Awasthi, P., Balcan, M. F., and K. Voevodski. (2014). "Local algorithms for interactive clustering." *ICML*. 550-558.

Bentham, J. 2009. "Punishment and Deterrence." In: *Principled Sentencing: Readings on Theory and Policy*, edited by von Hirsch, A., Ashworth, A., and J. Roberts. Oxford:Hart Publishing.

Calo R. 2015. "Robotics and the Lessons of Cyberlaw." *California Law Review*, no 103: 513.

Calo, R. 2016. "Robots in American Law". *University of Washington School of Law Research Paper*, no 99.

Chomsky, N. 1996. *Powers and Prospects.* London: Pluto Press.

Chopra, S., and L. White. 2011. *A Legal Theory for Autonomous Artificial Agents.* Michigan: The University of Michigan Press.

Citron, D. K. (2007). "Technological due process". *Washington University Law Review*, no 85: 1249-1313.

Citron, D. K., and F. A. Pasquale. 2014. "The scored society: due process for automated predictions". *Washington Law Review*, no *89*.

Crawford, K., and J. Schultz. 2014. "Big data and due process: Toward a framework to redress predictive privacy harms." *BCL Reviews*, no 55: 93.

Dick, Ph., 1968. *Do Androids Dream of Electric Sheep.* Bowling Green: Bowling Green.

Endsley, M. R., and D.B. Kaber. 1999. "Level of automation effects on performance, situation awareness and workload in a dynamic control task." *Ergonomics* 42(3): 462-492.

Fitts, P. 1951. "Human engineering for an effective air-navigation and traffic-control system." *Washington: National Research Council, Division of Anthropology and Psychology.*

Floridi, L. 2005. "Consciousness, agents and the knowledge game". *Minds and machines* 15(3–4): 415– 444.

Franklin, S. 1995. *Artificial Minds*. Boston, MA: MIT Press.

Gray, J. 1921. *The Nature and Sources of the Law*. London: Macmillan.

Haugeland, J. 1985. *Artificial Intelligence: The Very Idea*. Cambridge, Mass.: MIT Press.

Hildenbrandt, M. 2015. Smart Technologies and the End(s) of Law. Cheltenham, UK; Northampton, MA: Edward Elgar.

Holzinger, A. 2016. "Interactive machine learning for health informatics: when do we need the human-in-the-loop?" Brain Informatics, no 3:119-131.

Hull, L. 2013. "Doing the school run just got easier! Nissan unveils new car that can drive itself on short journeys" "Doing the school run just got easier! Nissan unveils new car that can drive itself on short journeys". Daily Mail (London). Retrieved 14 February 2016.

Copeland, J. 2014. *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life.* Oxford: Oxford University Press.

Jones, M. 2015. "Ironies of Automation Law: Tying Policy Knots with Fair Automation Practices Principles." *Vand. J. Ent. & Tech. L.*, no 18: 77.

Kaeb, C. 2015. "Putting the 'Corporate' back into corporate personhood." *Northwestern Journal of International Law and Business* 35(3).

Kerr I. 2001. "Ensuring the success of contract formation in agent mediated electronic commerce." *Electronic Commerce Research* 1: 183–202.

Kumar, D., Rai, C. S., and S. Kumar, S. 2010. "Analysis of unsupervised learning techniques for face recognition." *International Journal of Imaging Systems and Technology*, 20(3): 261-267.

Levinson, J., and S. Thrun. 2014. "Unsupervised calibration for multi-beam lasers." In: *Experimental Robotics*. Berlin: Springer.

Levinson, J., Askeland, J., Becker, J., Dolson, J., Held, D., Kammel, S., and M. Sokolsky. 2011. "Towards fully autonomous driving: Systems and algorithms." In: *Intelligent Vehicles Symposium*, IEEE.

Minsky, M. 1988. *The Society of Mind"* London: Pan Books.

Microsoft's Official Blog.2016. "Learning from Tay's Introduction." http://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., and S. Petersen. 2015. "Human-level control through deep reinforcement learning." *Nature*, *518* (7540), 529-533.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of machine learning*. MIT press.

Moris, E. 2007. "From Horse Power to Horsepower." *Access Magazine* 1(30).

Nakajima, C., Pontil, M., Heisele, B., and T. Poggio. (2000). "People recognition in image sequences by supervised learning." *MIT Report*. Cambridge MA, Center for Biological and Computational Learning.

National Highway Traffic Safety Administration.2008. "National Motor Vehicle Crash Causation Survey: Report to Congress." http://www.nrd-nhtsa.dot/gov/pubs/811059.pdf

Parasuraman R, 1997. "Humans and Automation: Use, Misuse, Disuse, Abuse", *Human Factors 37 (2):230-253*

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Perez, S (2016, March), "Microsoft Silences its new A.I. bot Tay, after Twitter users teach it Racism". https://techcrunch.com/2016/03/24/microsoft-silences-its-new-a-i-bot-tay-after-twitter-users-teach-it-racism/ (retrieved 12/08/2016)

Riley, V. 1989. "A general model of mixed-initiative human-machine systems." *Proceedings of the Human Factors Society*, no 33: 124-128.

Rosen, R. 2012. "Google's Self-Driving Cars: 300,000 Miles Logged, Not a Single Accident Under Computer Control." *The Atlantic.*

Russell, S J. and P. Norvig. 2003. *Artificial Intelligence: A Modern Approach.* New Jersey: Prentice Hall.

SAE International. 2014. *Surface Vehicle Information Report, J3016: Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems.*

Sandvig, C., et al. "Auditing algorithms: Research methods for detecting discrimination on internet platforms." 2014. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry.*

Scherer, M. 2016. "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies" *Harvard Journal of Law & Tech*, Vol 29 (2):354,359.

Sheridan, T. 1970. "On how often the supervisor should sample." *IEEE Transactions on Systems Science and Cybernetics.* SSC-6: 140-145.

Sheridan, T. (2000). "Function allocation: algorithm, alchemy of apostasy?" *International Journal of Human Computer Studies* 5(2):205.

Sheridan, T. B., and W.L. Verplank. 1978. *Human and computer control of undersea teleoperators.* Arlington: Office of Naval Research.

Silberg, G., Manassa, M., Everhart, K., Subramanian, D., Corley, M., Fraser, H., and V. Sinha. 2013. "Self-Driving Cars: Are We Ready." White paper *KPMG.*

Spitzer, C R.1987. Digital Avionics Systems, Englewood Cliffs, NJ Prentice Hall.

Solum, L. B. 1992. "Legal personhood for artificial intelligence." *North Carolina Law Review*, no 70: 1231.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1, No. 1). Cambridge: MIT press.

Teubner, G. 2007. “Rights of Non Humans? Electronic Agents and Animals as New Actors in Politics and Law.”In: *Lecture delivered on 17th January 2007 – Max Weber Programme, European University Institute.*

Thompson J M. 1994. “Medical Decision Making and Automation”, in Mouloua, M. and Parasuraman, R. (eds) *Human Performance in Automated Systems: Current Research and Trends*, Hillsdale, NJ:Erlbaum.

Titiriga, R. 2016. “Autonomy of Military Robots: Assessing the Technical and Legal ('Jus in Bello') Thresholds.” *The John Marshall Journal of Information Technology & Privacy Law* 32(2): 57-88.

Trimble, T. E., Bishop, R., Morgan, J. F., and M. Blanco. 2014. *Human factors evaluation of level 2 and level 3 automated driving concepts: Past research, state of automation technology, and emerging system concepts.* Report No. DOT HS 812 043. Washington, DC: National Highway Traffic Safety Administration.

Turing, A. 1950. “Computing Machinery and Intelligence”, Mind, New Series, 59 (236): 433-460.

Urmson, C., Anhalt, J., Bagnell, D., Baker, C., Bittner, R., Clark, M. N., and M. Gittleman. 2008. “Autonomous driving in urban environments: Boss and the urban challenge.” *Journal of Field Robotics* 25(8): 425-466.

Van Hasselt, H., Guez, A., and D. Silver. 2015. Deep reinforcement learning with double Q-learning. *CoRR, abs/1509.06461*.

Wei, J., Snider, J. M., Kim, J., Dolan, J. M., Rajkumar, R., & Litkouhi, B. 2013. Towards a viable autonomous driving research platform. In: *Intelligent Vehicles Symposium*.

Williams A., Oliver C., Aumer K. and Ch. Meyers. 2016. “Racial Microaggressions and perceptions of Internet memes.” *Computers in Human Behavior* 63: 424-432.

Wilson, A., Fern, A., Ray, S., and P. Tadepalli. 2007. “Multi-task reinforcement learning: a hierarchical Bayesian approach.” In: *Proceedings of the 24th international conference on Machine learning*.

Wittenberg, S. 2016. “Automated Vehicles: Strict Products Liability, Negligence Liability and Proliferation”, Illinois Business L J

Zarsky, T. 2016. “The trouble with algorithmic decisions an analytic road map to examine efficiency and fairness in automated and opaque decision making.” *Science, Technology & Human Values* 41(1): 118-132.

Zhu, J., Montemerlo, M. S., Urmson, C. P., and A. Chatham. 2014. *U.S. Patent No. 8,874,372*. Washington, DC: U.S. Patent and Trademark Office.