MIT Computational Law Report

When Humans Judge Other Humans Using Machines

Renita Murimi

Published on: May 14, 2021

ABSTRACT

> *The role of intentions in determining important outcomes has been extensively studied. Uncovering intent remains a difficult and, at times, elusive quest where the outcomes have the potential for life-altering consequences. When machines are harnessed to aid in the decision-making process, the underlying algorithms, and thereby the inherent algorithmic biases, play a role in the outcomes of the process. This paper develops an analytical framework to study the role of human and machine perceptions of intentions and the resulting outcomes in the decision-making process. The work in this paper raises important questions about the impact of utilizing machines in judgments, and how machines are poised to deliver outcomes that are different from those of humans without adequate oversight and accountability.*

Introduction

Seneca remarked that the power of a gift lies in the intention of the gift-giver, rather than in the gift itself. Over time, philosophers have extensively pondered on the role of intentions in the formation of beliefs and the execution of those intentions in the form of actions or outcomes. The role of intentions in influencing outcomes was described by the economist Milton Friedman in his famous words, “One of the great mistakes is to judge policies and programs by their intentions rather than their results.”

Law has deeply sought to separate intentions from outcomes, and this quest forms the cornerstone of criminal law. Deciphering intent has engendered robust discussions, summoning related streams of thought in sociology, philosophy and law. As of this writing, a quick search of Supreme Court of the United States case law reveals more than five hundred instances of opinions that aim to uncover the intent underlying the case. An example of the role of intent influencing the outcome is exemplified in *Dean v. United States* (Dean, 2009) where the issue aimed to answer if a federal law imposing a ten-year minimum additional sentence on a defendant who fired a gun during a violent crime applied even if the firing was an accident. In 2004, Christopher Dean and his brother-in-law, Ricardo Lopez, entered the Rome location of AmSouth’s bank in Georgia with a pistol. While attempting to take money from the multiple teller stations, the gun was discharged. In this case, a 7-2 decision was delivered by the Supreme Court stating that no proof of intent was required. Dean was convicted of conspiring to commit a robbery and discharging a firearm during the armed robbery, for which he

was sentenced to ten years in prison under Title 18 U. S. C. §924(c)(1)(A). The two opening sentences of this opinion delivered by Chief Justice Roberts read, “Accidents happen. Sometimes they happen to individuals committing crimes with loaded guns”.

In arriving at this decision in 2009, it is reasonable to assume that technology helped in every step of the way, starting from the moment that the police arrested Christopher Dean. Various forms of technology involving devices, software and networked environments assisted the efforts of legal professionals along the process. It is also reasonable to assume that the technology did not influence the outcomes of the courts, but rather merely served as a tool for archiving, querying and recording information.

However, technology might not be a benign tool for much longer. Five years after Dean was convicted and sentenced, DoNotPay, an AI-powered software developed by Josh Browder was used to overturn a parking ticket (Nunez, 2017), (Simshaw, 2018). The driver answered a few questions posed by DoNotPay, and without interacting with any humans, DoNotPay filled out the forms, and filed them free of charge. Over the next three years, DoNotPay was able to dispute 375,000 parking tickets and saved people more than 9 million dollars. Robo-lawyering had entered a new dimension. Since then, the company has morphed into a portal that aids in various kinds of disputes, including canceling services or subscriptions, issuing refunds, canceling free trials and filtering spam. Powered by IBM’s Watson, DoNotPay calls itself the “world’s first robot lawyer” that “fights corporations, beats bureaucracy and sues anyone at the press of a button”.

It is not hard to observe the outcomes of DoNotPay. However, it is more difficult to understand its intentions. Ethicists will point to the dichotomy between the intentions of the robot versus the intentions of its creators. Indeed, the work of Hidalgo *et al.* (2021) poses a thought-provoking question: how do humans judge machines? Hidalgo *et al.* analyzed this question from a positive approach rather than a normative approach by using extensive experimental data, and found that humans judge humans by their intentions, but judge machines by their outcomes. This finding has important implications for accidents – a machine committing an accident is judged more harshly than a human committing an accident.

A different example of “intentions gone awry” is found in Inspirobot. Inspirobot is a website that generates inspirational quotes superimposed on pictures, resembling the inspirational quotes found on social media feeds. However, instead of parsing through databases of quotes and pulling up a random quote, or a quote related to a thematic keyword, Inspirobot is an AI-powered quote generator. The quotes produced by Inspirobot can be comical. One of the quotes generated by Inspirobot – “Seek success,

but prepare for vegetables" – is an example of an inspirational quote gone comically wrong. Ironically, another quote from Inspirobot said, "Will humans ever be able to handle intelligent discourse?"

In this paper, we extend this premise of intentions and outcomes in the judging process to an analytical form. Specifically, we study the role of three factors: perceptions of intentions, perceptions of outcomes, and importance of these perceptions to an agent that influences the outcomes of the judging process. To highlight the contrast between scenarios that involve machines and those that do not, we first propose a simple model where humans judge humans. Next, we extend this model to study how humans judge humans with the aid of a machine. In each case, we provide tractable expressions for the outcomes, and comment on the model's applications, limitations and directions for future work.

Related Work

The confluence of technology and law has been studied since ancient times. One of the earliest references to a legal code was found in early Mesopotamian laws. The Code of Urukagina exemplified the need to achieve more freedom and equality, creating the first known recorded instance of the word "freedom" as "ama-gi" (Finegan, 2019). Aristotelian teachings gave rise to the Rule of Law, which set forth the idea that all people and all institutions were to be held accountable to the same set of laws. As technology advanced, the tools advanced too, ranging from scrolls and writing instruments to modern day digitally networked environments. Thus, law has always benefited from the technology of the times, mainly because both law and technology are essential components of a well-functioning society.

However, modern day technology is on a trajectory to serve society as more than just a tool; without enough caution, it is on a path to become the judge and arbiter. The algorithms that power the software and devices of our time are instrumental in determining access to education, eligibility for employment, housing, and even incarceration (O'Neill, 2016). The sophistication of these tools varies, and trends in data mining and AI have enabled applications that complement human counterparts in sifting through troves of data and uncovering patterns. That might be where the prowess of present day technology in accomplishing higher-order tasks such as reasoning ends. Frequently employed in legal decision making, the work of Brewer (1996) describes the process of reasoning by analogy in a three-step process. In the first step, inference from chosen examples is used to point to a rule that eliminates doubt. In the second step, the discovered rule is confirmed or discarded by dynamic

reflection. In the final step, the confirmed rule is applied to the case under consideration. When computers are called upon to participate in this process of reasoning by analogy, they become active participants in the inference-confirmation-application process. However, while computers are capable of broadly recognizing patterns, they lack an understanding of what these patterns stand for.

About three decades ago, Rissland (1990) took a long view of the fields of AI and law, and enumerated several goals that an ideal AI and law program would achieve. Among other things, this ideal program would be capable of reasoning, modeling and understanding knowledge, intents and beliefs. While AI has advanced a great deal since the time of that writing, we are still far away from actualizing a program with these capabilities. Still, significant strides have been made. The work of Love and Genesereth (2005) posits that the application of logic-based techniques to law, which includes the development of software tools, business-process languages and the use of techniques from behavioral modeling and state machines, could be used to create robust applications for computational law. Additional work from Genesereth (2015) contends that computational law can help simplify the law and make it accessible. As an example, he illustrates the use of Intuit's TurboTax as an elementary computational law system that helps users file their taxes through an intuitive interface that does not require the user to possess any understanding of tax law. However, the author points out important limitations concerning the extent of the computational systems' reach. The computational system can only understand and execute the legal semantics that it can parse through logical arguments; it cannot read between the lines. As such, he makes a distinction between the kinds of jurisprudence that computational law is well-served to benefit. Computational law, by design, can aid in the legal formalism school of jurisprudence which treats laws as definitive. On the other hand, computational law cannot fully assist in legal realism, where outcomes are decided on a case-by-case basis. Thus, while computational law is of potential interest in civil law, the lack of innovation afforded by computational law makes it of scarce use in common law settings.

This implies that we have more to accomplish when it comes to applications of technology to law. As an example, consider the role of expert opinions. The work of Walton and Gordon (2005) addresses six critical questions regarding arguments from expert opinion. These questions pertain to the credibility of the expert, their expertise in a particular domain, the implications of the expert's opinion, reliability of the expert, consistency of the expert's opinion and the role of evidence in the expert's opinion. These six foundational questions provide a framework to address weaknesses in an

argument, and help to formulate the groundwork for the role of the machine in computational law. Additional work by Verheij (2003) distinguishes four different roles of critical questions in an argument – premises, exceptional situations, conditions for use, and other possible arguments that could lead to the same conclusion. Yulil (2019) further highlights the role of expert systems in applying AI to legal systems. Online dispute resolution systems, downloadable templates for wills and legal agreements, contracts and trading exemplify how algorithmic abilities can enhance the application of law to everyday transactions in various domains. Further, the author explains how law and computation complement each other to create order through logic.

Explainability represents a key milestone in the learning process for both humans and AI algorithms. For example, if one were to ask a smart speaker for the sum of 2+3, it would respond with 5. But when asked to explain why the sum was 5, it would respond with a vague answer such as “I did not understand the question” or “Hmmm…”. We rely on humans to act, execute and explain their work. When machines are employed to aid in key outcomes, it is crucial to understand how the machine arrived at a decision. For example, the use of specific algorithms, assumptions, inputs and training datasets would help to clarify how the machine worked through the process to produce the output. The work of Hildebrandt (2020) explains additional issues surrounding explainability of decisions. Here, the author argues that law forces judges to explain their decisions by constraining their decision spaces. Machines that are employed to help make decisions would also have to adhere to the framework of explainability, so that the same level of accountability can be attributed to the machines used in computational law. McLaren (2006) provides an overview of computational models used in ethical reasoning, and suggests that such models are best used as teaching aides due to their use of simplifying assumptions that do not translate well to complex real-world cases.

The current state of computational law can be studied in terms of its applications. Linna (2019) makes the distinction of AI being used for two broad categories of legal work. The first category pertains to the ethics, regulations and laws that apply to technology, whereas the second category pertains to the ways in which technology offers tools to improve legal services. The author raises questions related to the intersection of code and law, and suggests that cross-disciplinary collaboration between lawyers and technologists is required to address the challenges of our complex digital society. Further work of Lettieri *et al.* (2018) presents an overview of legal analytical platforms. A significant driver of the trends in legal technology is the set of tools associated with big data that allows for discovering facts and precedents,

thus uncovering trends and patterns in cases. AI tools take this a step further by building intuitive discovery machines that provide relevant cases with degrees of similarity, as well as prediction machines that are harnessed in applications such as crime prediction and criminal network analysis.

Smith (2007) describes why any form of technology will fail to outperform human intelligence. Borrowing upon work from Bench-Capon (1990), Smith explains how the four components of human intelligence – consciousness, self, perception, and language – cannot be coded in a machine in a way that matches the faculties of the human. Sunstein (2001) provides evidence of how computers are unable to reason analogically like humans, and therefore are not capable of providing effective legal outcomes in decision making . The author reasons that AI exists in two forms – strong and weak. The weak version offers improved tools for document discovery, database queries, analysis of patterns and trends. The strong version of AI, on the other hand, is required to perform legal reasoning on par with humans. The author contends that current AI technology, while offering AI tools to deliver services in the weak version, is incapable of functioning at a level required by the strong version since computers cannot make evaluative arguments.

The work in this paper aims to study the role of the machine in judgements. Specifically, we study two models – one where humans judge humans without machines, and the other where humans judge humans with the aid of machines. In each case, we provide analytical expressions for the outcomes of the judgement, considering the roles of perception and bias. The next section describes the model.

The Model

Assume a preliminary model as shown in Figure 1. An agent, represented either as a judge, the accused, or a machine, is assumed to have intentions and outcomes. We refrain from using the term “moral agent” since we are not assuming that the machine has the ability to discern right from wrong. Neither do we attribute moral status to the machine, which confers on them the status accorded to a human agent in the judging process.

Figure 1 - A preliminary model of intentions and outcomes for an agent.

We extend the preliminary model shown in Figure 1 to incorporate the process of a human judge who is judging an accused party. Figure 2 shows this extended model with two agents - H_d, the human judge and H_e, the accused. For the remainder of this paper, we will assume that the judge and the accused are both human (although this may not be the case, in the future), and so we will omit the "human" qualifier. The agent H_d has intentions I_d and delivers outcomes of the judging process as O_d. Our goal in this paper is to propose an analytical expression for O_d that considers the role of perceptions and biases in human and machine agents involved in the judging process.

Let p_i denote the judge's perceptions of the accused's intentions, and let p_o denote the judge's perceptions of the accused's outcomes. Further, let α_i denote the weight assigned by the judge to the accused's intentions, and let α_o denote the weight assigned by the judge to the accused's outcomes. Here, $p_i,\ p_j \in \{0,1\}$, where a value of 0 signals that the judge's perception of the intentions or outcomes is completely opposite from the actual case, and a value of 1 denotes that the judge's perception is identical to what happened in the actual case. Further, we assume that the judge is cognizant of her own biases with a level σ, where $\sigma \in \{0,1\}$. A level of 0 indicates that the judge is unaware of her own biases in the intentions, and 1 indicates the judge is aware of her own intentional biases. Thus, σ, p_i and p_j offer room for the variability of the fidelity of evidence, the strength of arguments and the judge's own biases in proceeding to deliver judgement.

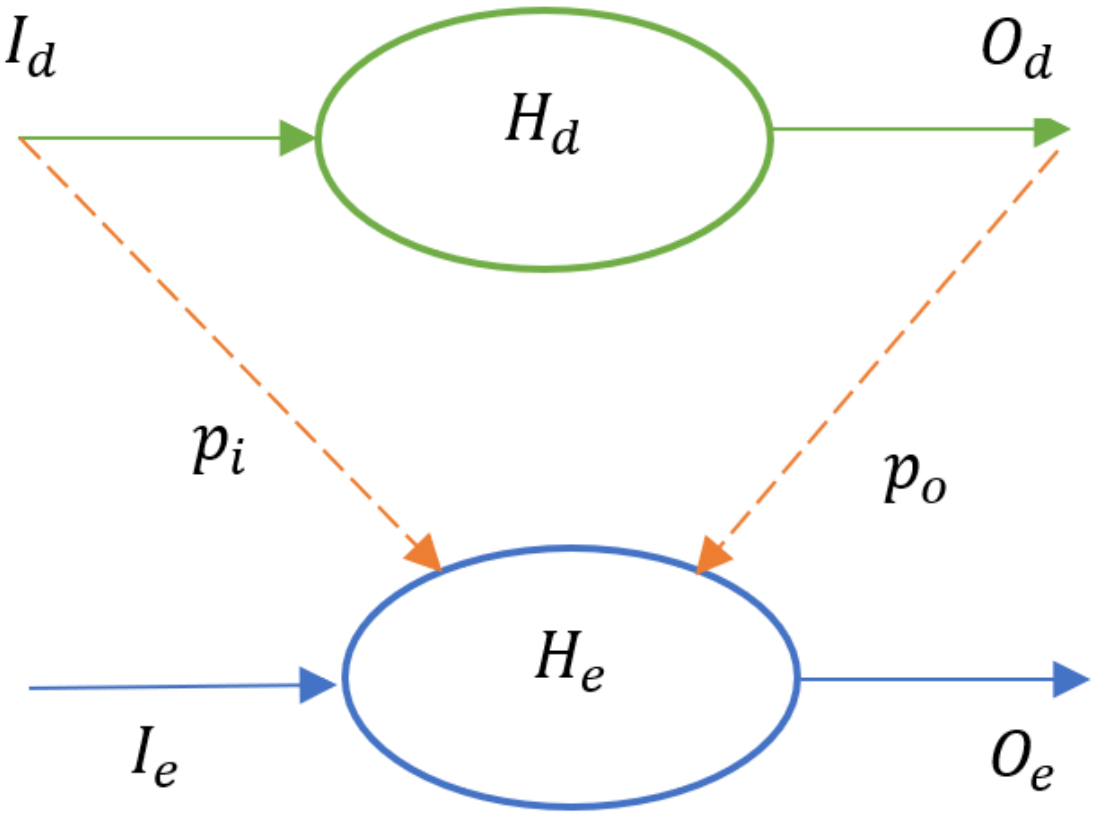

Figure 2 - A model for judge-accused interaction.

The judge's outcome, which refers to the outcome of the decision process, is a function of her own intentions, and her perception of the intentions and outcomes of the accused, as well as the weight assigned to these perceptions. The outcome of the judge's decision process is then given by equation (1) below, where the first term in parentheses on the right-hand side refers to the judge's perception and associated weights of the accused's intentions and outcomes. The second term on the right-hand side refers to the judge's own intentions concerning the case involving the accused.

$$O_d = (\alpha_i \; p_i \; I_e + \alpha_o \; p_o \; O_e) + \sigma \; I_d \quad (1)$$

Next, we extend this model further to include the role of the machine in the judging process. Figure 3 shows an additional agent – the machine, M. The judge, H_d, consults or uses the services of M in judging a case related to the accused, H_e. The machine M has outcomes o_m that are used by the judge in the decision process of the accused. The

machine is assumed to have its own perceptions of the accused's intentions (m_i) and the accused's outcomes (m_o). Similar to the judge, the machine also ascribes weights to the accused's intentions (γ_i) and the accused's outcomes (γ_j). The values m_i, m_o, γ_i, and γ_o lie on the $\{0, 1\}$ continuum representing varying machine perceptions and the weights assigned to the perceptions, and allows room for algorithmic biases, and varying quality of evidence and arguments, in the machine's computational processes.

The outcomes of the judge's decision process that has been aided by a machine are then given by:

$$O_m = (\alpha_i\, p_i\, I_e + \alpha_o\, p_o\, O_e) + \sigma\, I_d + (\gamma_i\, m_i\, I_e + \gamma_o\, m_o\, O_e)$$

Judge decision

Judge –Accused Relationship

Self

Machine –Accused relationship

(2)

Comparing equations (1) and (2) shows that the introduction of the machine alters the nature of the decision process. Depending on the level of inherent biases on the part of both the machine and the judge, existing flaws in the evidence or arguments can be exacerbated, potentially leading to outcomes that are unfair. In the next section, we present a simulation of equations (1) and (2) to illustrate the role of perceptions and biases in both the machine and the human in the outcomes of the judging process.

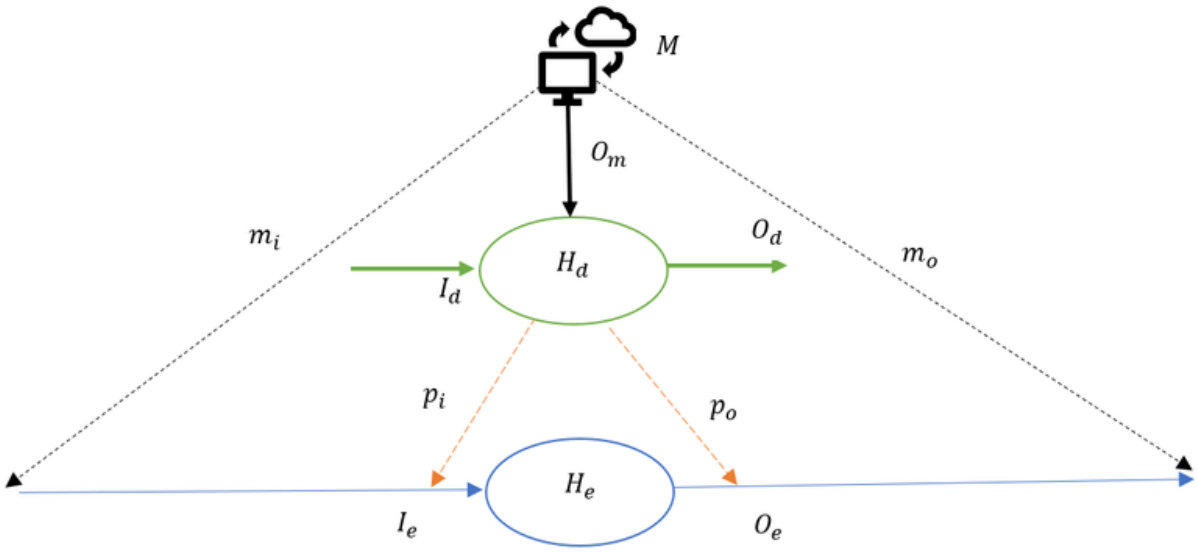

Figure 3: A model for judge-machine-accused interaction.

Limitations

The work in this paper proposed an analytical model that considers intentions, outcomes, perceptions and biases in both human and machine agents in the legal process. The model offers a mathematical expression for the outcomes of the legal process based on several simplistic assumptions. While these assumptions make for a tractable model, they also present avenues for refinement of the model to more closely emulate real-world environments. A few of the limitations are described below.

1. Solitary agents: Our model considered a single judge who judges the outcomes of a legal case concerning a single accused individual, with the aid of a single machine. In practice, however, multiple agents collaborate in the legal process, bringing forth a variety of perspectives, and thereby adding to a more thorough vetting of the facts and arguments. An analysis of such multi-agent environments would require the development of complex models with multivariate factors that contribute to the development of legal outcomes.
2. Static linear weights: In our model, we assume that the weights (α, γ, and σ) assigned to the intentions, outcomes and perceptions of both humans and machines are linear. The weights signify the importance attached to each of these variables in the legal process. Further, we assume that these variables are static. However, real-world environments are rarely linear or static. For example, in the process of determining an outcome, discovery of certain facts or the occurrence of particular instances might change the importance of a specific variable to the outcome.

Additionally, different legal environments might impose additional variables and weights to their specific cases. While our model provides a foundational mechanism to think about the impact of an agent's perception of the case, further work is required to develop application-specific versions of our model.

3. Modeling specific biases: The role of biases in judgements has been captured in our model through the perceptions of each agent and their associated weights. However, the broad category of biases can be analyzed further to address domain-specific or case-specific biases. Further, prevailing socio-economic conventions introduce additional nuances to the modeling of biases in the legal process. A deeper understanding of biases for both human and machine agents will help create more robust models that capture the richness of real environments.

Future Work

This paper proposed an analytical model to study the role of the machine in delivering judgements. Multiple applications of this model can be envisioned, including the implications of granting autonomy to the system, and the standardization in AI systems used in computational law.

Granting autonomy to the machine: Granting autonomy to the machine could allow it to choose outcomes outside of the narrow decision space afforded to the machine. While this might not be a desirable outcome, granting autonomy allows the machine to make "mistakes" on its path to learning the right rules and assumptions. Neural networks already use a simplified version of autonomy, allowing them to reward correct decision paths and penalize wrong outcomes. The ability to choose from a wider decision space allows the machine to exhibit a level of creativity, albeit while promoting indeterminacy. However, it is important to establish upper bounds for indeterminacy so that the machine outcomes remain within acceptable bounds of freedom. Such an autonomous machine would create continuously changing legal systems that adapt to the technology, and ultimately the society. For example, legal systems that are based on precedent rely upon the most recent instance of the rule. An autonomous machine that dynamically changes its knowledge base of rules and precedents would be able to exercise greater innovation in interpreting the law, while affording researchers and developers the ability to study the impact of the autonomy and algorithmic bias on the outcomes of the decision process.

Standardization: Various technologies such as electronic discovery software, drafting software, and case analysis software, exist to provide legal professionals both general and specific solutions. The use of AI algorithms in computational law, however, calls for

standardization. Law in its current form, is intentionally vague and open to interpretation. Without standardizing AI for its applications in computational law, algorithms may acquire alternative legal premises stemming from alternative legal ideologies, which can lead to different outcomes (Sartor, 1993).

Conclusions

This paper addressed the role of machines in the legal process with the help of an analytical model that considered three variables: intentions, outcomes and perceptions of the agent involved in the legal process. The role of machines in furthering the field of computational law is a *sine qua non*; still, it behooves us to think carefully about the nature of the machine's involvement and its role on the outcome of the judgement process. As machines make their way into an increasing number of our environments, the law that regulates these environments will inevitably use this technology for efficiency. Models such as the one presented in this paper provide room for discussion about the impact of technology's involvement in our lives and that of our society.

References

Bench-Capon, T. J. M. (1990). Knowledge representation: An approach to Artificial Intelligence. Vol. 32. Elsevier.

Brewer, S. (1996). Exemplary reasoning: Semantics, pragmatics, and the rational force of legal argument by analogy. *Harv. L. Rev.*, 923-1028.

Dean v. *United States,* 556 U. S. 568 (2009).

Finegan, J, (2019). Archaeology of the Ancient Middle East. Routledge.

Genesereth, M. (2015). Computational Law. *The Cop in the Backseat*. White Paper, CodeX—The Stanford Center for Legal Informatics.

Hidalgo, C., Orghiain, D., Canals, J.A., de Almeida, F., Martin, N. (2021). How Humans Judge Machines. MIT Press.

Hildebrandt, M. (2020). A philosophy of technology for computational law. *The Philosophical Foundations of Information Technology Law*, Oxford University Press.

Lettieri, N., Altamura, A., Giugno, R., Guarino, A., Malandrino, D., Pulvirenti, A., & Zaccagnino, R. (2018). Ex machina: Analytical platforms, law and the challenges of computational legal science. *Future Internet*, *10*(5), 37.

Linna Jr, D. W. (2019). The Future of Law and Computational Technologies: Two Sides of the Same Coin, *MIT Computational Law Report*, Release 1.0.

Love, N., & Genesereth, M. (2005, June). Computational law. *Proceedings of the 10th international conference on Artificial intelligence and law* (pp. 205-209).

McLaren, B. M. (2006). Computational models of ethical reasoning: Challenges, initial steps, and future directions. *IEEE intelligent systems*, *21*(4), 29-37.

Nunez, C. (2017). Artificial intelligence and legal ethics: Whether AI Lawyers can make ethical decisions. *Tul. J. Tech. & Intell. Prop.*, *20*, 189.

O'Neill, C. (2016) Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Broadway Books.

Rissland, E. L. (1990). Artificial intelligence and law: Stepping stones to a model of legal reasoning. *The Yale L.J.*, *99*(8), 1957-1981.

Sartor, G. (1993, August). A simple computational model for nonmonotonic and adversarial legal reasoning. *Proceedings of the 4th international conference on Artificial intelligence and law* (pp. 192-201).

Simshaw, D. (2018). Ethical issues in robo-lawyering: The need for guidance on developing and using artificial intelligence in the practice of law. *Hastings L.J.*, *70*, 173.

Smith, J. C. (1997). Machine Intelligence and Legal Reasoning. *Chi.-Kent L. Rev.*, *73*, 277.

Sunstein, C. R. (2001). Of artificial intelligence and legal reasoning. *U. Chi. L. Sch. Roundtable*, *8*, 29.

Verheij, B. (2003). Dialectical argumentation with argumentation schemes: An approach to legal logic. Artificial Intelligence and Law, 11(2-3):167–195, 2003.

Walton, D., & Gordon, T. F. (2005). Critical questions in computational models of legal argument. *Argumentation in artificial intelligence and law*, 103.

Yuill, S. (2019). Section Editorial: Critical Approaches to Computational Law. *Computational Culture*, (7).