The Neglected Dualism Of Artificial Moral Agency And Artificial Legal Reasoning In AI For Social Good

Dr. Lance B. Eliot
Chief AI Scientist, Techbruim; Fellow, CodeX: Stanford Center for Legal Informatics
Stanford, California, USA

Abstract

A neglected dualism is occurring in AI for Social Good involving the lack of encompassing both the role of artificial moral agency and artificial legal reasoning in advanced AI systems. Efforts by AI researchers and AI developers have tended to focus on how to craft and embed artificial moral agents to guide moral decision making when an AI system is operating in the field but have not also focused on and coupled the use of artificial legal reasoning capabilities, which is equally necessary for robust moral and legal outcomes. This paper addresses this problematic neglect and offers insights to overcome a substantive prevailing weakness and vulnerability.

1 Introduction

A key question in AI for Social Good is how to ensure that AI systems operating in the field are able to make needed and appropriate moral choices. As society becomes increasingly dependent on AI, there is a widening concern that AI systems might at times not conform to applicable moral norms. By crafting and embedding Artificial Moral Agents, specialized components within an AI system, it is anticipated that a real-time capacity will enable the enactment and abidance of moral conventions [Cane, 2012; Huang, 2019; Misselhorn, 2019].

Even if Artificial Moral Agents can be suitably devised, there is a missing and crucial element that is currently being neglected, namely the dual role of morality and law. As pointed out by Shiell [1987], since at least the days of Plato there has been a struggle over the relationship between morality and law, and even still today debates about the nature and extent of how moral decisions intertwine with the law are persistent and contentious.

Overall, morality and law are generally viewed as inextricably bound to each other in some fashion and must be considered in concert, though their mutual interaction and dependencies are still an open-ended matter

As such, additional attention to Artificial Legal Reasoners (ALR) is needed, providing the other missing or underplayed element for achieving the requisite morality and law dualism. The inclusion of artificial legal reasoning provides the possibility of including a cooperating specialized ALR component within an AI system that offers a real-time capability for rendering legal-based decisions or awareness about the law [Eliot, 2019; Genesereth, 2009; Surden, 2019].

By considering the simultaneous use and deployment of components for both Artificial Moral Agent capabilities and Artificial Legal Reasoner capabilities, working in a coordinated and communicative manner, the moral agency of AI systems is likely to be more well-rounded and balanced by prevailing laws. This aspiration though is not as readily fulfilled as might seem at first glance since there are inherent tensions between morality and law, which will be further exemplified and revealed in a collaborative effort to have such automated agents work in conjunction with each other.

The remainder of this paper describes the possibilities and problems that will be encountered in seeking to achieve the proffered dualism. In addition, insights about ways to cope with the consequent hurdles and difficulties are laid out as a research agenda for those pursuing the development of AI for Social Good.

2 Dualism Tensions

As stated by Bickenbach [1989], “the relationship between morality and law is one of the more enduring problematics of jurisprudence.” Those legal scholars in the domain of natural law argue that morality is the source of laws and serves as the binding power of laws. In contrast, scholars considered in the legal positivism domain are likely to contend that morality is categorically separate and distinct from law.

In a legal realism context, Kagan [1998] points out that morals and laws ultimately exhibit tensions between each other, and emphasizes that “the law may permit some particular act, even though that act is immoral; and the law may forbid an act, even though that act is morally permissible, or even morally required.”

Morality is sometimes viewed as existing within the shadow of law [Mnookin and Kornhauser, 1979]. In that sense, a moral facet would be considered outside of the law and yet within reach of the law. Some view that morality is actually the halo of law, such as Regan [1987] indicating that law is imbued with moral sustenance, even if at times there might not be a moral obligation to abide by the law.

Consider a given body of moral tenets as represented by a designation M and a body of laws as represented by the use of a designation L. Envision these as circles or ovals for which there is a Venn diagram depicting them as overlapping, whereby the designator O represents their overlap:

$$\boldsymbol{O = M \cap L} \qquad \textbf{(1)}$$

Define M' and L' as follows:

$$\boldsymbol{M' = M - (M \cap L)} \qquad \textbf{(2)}$$

$$\mathbf{L' = L} - \boldsymbol{(M \cap L)} \qquad \textbf{(3)}$$

If and only if M′ is fully outside of the legal realm and not subject to the law, it will be assumed that any instance of an Artificial Moral Agent rendering such a moral choice within the scope of M′ could ergo proceed unabated without any needed consultation with an Artificial Legal Reasoner.

Likewise, if and only if L′ is fully outside of the moral realm and not subject to a moral underpinning, it will be assumed that any instance of an Artificial Legal Reasoner rendering such a legal indication could ergo proceed unabated without any needed consultation with an Artificial Moral Agent.

The O presents the key challenge for the dualistic nature of the two.

Divide O into those instances for which there is an agreement between the given M and L, which will be designated as A, and those instances for which there is disagreement which is to be designated as D.

$$\mathbf{O = A + D} \qquad \textbf{(4)}$$

In the case of A, the presumption is that since the M and L are in concordance on such instances, the Artificial Moral Agent and the Artificial Legal Reasoner have no conflict and thus either one can prevail in such a use case.

We are then left with the class D of instances, providing the essence of conflict that needs to be resolved or otherwise conveyed. Some form of automated conflict resolution will be required to contend with body D. In the study of law, there are considered "hard" cases that are less amenable to everyday legal reasoning [Hage *et al*, 1993].

Class D can be characterized as being composed of some combination of hard case instances, designated as D* and those that are non-hard, depicted as D° (do not conflate the notion of non-hard with a meaning of being easy or simple, since the non-hard cases can also pose quite arduous and complicated challenges).

For D then:

$$\mathbf{D = D^{*} + D^{\circ}} \qquad \textbf{(5)}$$

Algorithmically, an AI for Social Good system that makes use of an Artificial Moral Agent should correspondingly include an Artificial Legal Reasoner, and for which at any invoking of either one, the other should also be invoked, and the AI then would make a comparison of the results so rendered by each respectively.

For M′, the AI proceeds with the result of the Artificial Moral Agent, and for the L′ the AI proceeds with the result of the Artificial Legal Reasoner.

In the O, the AI can choose from either one in the instance of A, while for those that are D the AI would undertake a conflict resolution process (as discussed next). It is likely that the D* will require an elaborate effort by the conflict resolution process, while the D° will be predominantly less protracted.

3 Conflict Resolution

In this effort of the prescribed dualism, an a priori approach will need to be established for resolving the potential conflicts between the results tendered by the Artificial Moral Agent and those of the Artificial Legal Reasoner.

Undertaking a satisfactory conflict resolution of this sort is not straightforward, as this salient remark in *The Law* by Frederic Bastiat [1850] illuminates: "When law and morality contradict each other, the citizen has the cruel alternative of either losing his moral sense or losing his respect for the law."

An outline of the potential approaches to the morality and law conflicts in an AI for Social Good system is indicated next, numbering each approach as Cn, and for which the n is merely for reference purposes and not to suggest priority or sequence of which approach is valued over another.

3.1 Approach C1: Morality Prevails Over Law

In Hart [1961], morality is depicted as the ultimate standard for assessing human behavior and thus the law is considered second-best, namely that moral reasoning eclipses any legal reasoning.

As per Goparaju Ramachandra Rao [1980] in *I Learn:* "Whenever legality clashes with morality, legality should be opposed, and morality should be upheld."

In that case, for a conflict resolution approach labeled as C1, for those instances in the realm of class D, the M will always prevail, regardless of being either D° or D*.

This does not obviate the need for the use of an Artificial Legal Reasoner since there is still the class of L′ to be dealt with. It does though significantly reduce the run-time effort since anything other than L′ is transferred over to the Artificial Moral Agent to render a final decision.

3.2 Approach C2: Law Prevails Over Morality

In approach C2, the law is considered to prevail over the side of morality, and therefore any instances in class D are to be decided by the Artificial Legal Reasoner. This might be likened as a variant of the Rule of Law as exemplified by Albert Dicey's [1885] quote: "With us no man is above the law and every man, whatever be his rank or condition, is subject to the ordinary law of the realm and amenable to the jurisdiction of the ordinary tribunals." Thus, even if the Artificial Moral Agent has deemed that the instance is not morally aligned, nonetheless, the law shall prevail.

3.3 Approach C3: Determining Which Prevails

In approach C3, it is presumed that both the Artificial Moral Agent and the Artificial Legal Reasoner have a bona fide basis for why their respective rendered decisions are in conflict with each other and that the rules of C1 and C2 are not applicable.

Therefore, if a choice is to be made between the two, there must be some means to make such a choice.

Quite a number of algorithmic avenues could be utilized. Some are mentioned herein, each of which has tradeoffs and the particular class of AI for Social Good system will be a determiner in which such avenue is warranted. Also, this is by no means an exhaustive list and merely indicative of representative ways that conflicts might be resolved.

For example, there could be a weighting scheme that provides weights associated with rendered choices. This potentially introduces the use of uncertainty and probabilities into the Artificial Moral Agent and the Artificial Legal Reasoner, which is an advanced variant that some believe is needed in any case, regardless of this specific use for conflict resolution [Bench-Capon, 2020; Brandao *et al*, 2020].

Another approach would be to utilize an Arguing Machines methodology [Eliot, 2018; Fridman, 2017], consisting of the two components engaging in a dialogue or argument with each other, trying to convince the other that their choice ought to prevail on their own side of the matter.

Whichever avenue is used, there are key considerations to be attended to. If the run-time execution of attempting to settle the open conflict is onerous, the impact of the conflict resolution can undermine the overarching actions of the AI for Social Good system. A delay in responsiveness might be more than simply aggravating or inconvenient since the AI system might be immersed in a real-time activity that entails life-or-death decisions of an extremely timely nature (such as in the case of autonomous systems, including self-driving cars [Eliot, 2016; Huang, 2019]).

A sense of AI self-awareness [Parasuraman et al, 2000] is required as part of the conflict resolution process, namely that the AI must be keeping tabs on the conflict resolution and have some means of ascertaining that either the process is taking too long for the matter at hand or that the process has potentially become indeterminate or intractable, and needs to be interrupted or halted [Eliot, 2017]..

3.4 Approach C4: Neither Prevails

In approach C4, the question arises as to what the AI should do if the conflict between the Artificial Moral Agent and the Artificial Legal Reasoner cannot be otherwise resolved. Assume that C1 is not applicable, nor C2, and nor has C3 reached a resolution.

A final catchall that perhaps randomly picks between the two is conceivable, though likely unsatisfying in many respects, or apply a method that tries to assess whether the choice of one is somehow preferred over the other. Researchers such as Bench-Capon [2020] have identified options such as consequentialism might be used (the impacts of the activities chosen), or deontologically chosen (a worth associated with the act, irrespective of the consequences), or even a means abiding by Maslow's [Maslow, 1943] hierarchy of needs (a selection based on the option fulfilling the highest basic human need).

Another possibility is to seek a resolution from the end-user of the AI for Social Good system, including possibly offering an explanation associated with the impasse (using XAI, as described by Waltl and Vogl [2018]). Seeking such input from the end-user could be problematic in many ways, including allowing a selection preference with unintended consequences or other untoward possibilities and should carefully be assessed as to its utility [Freedman *et al*, 2020].

4 Discussion

This paper provides insights into the neglected dualism of Artificial Moral Agents and Artificial Legal Reasoners and provides an indication as to the value of the moral-and-law dualism, along with offering ways to encompass that dualism. It is hoped that this study will spur additional research into an emerging area that is only yet being explored and will likely become increasingly crucial for the expanding and widespread adoption of AI for Social Good.

About The Author

Prior to his present position, Dr. Eliot served as a professor at the University of Southern California (USC) and headed a pioneering AI lab, his AI textbooks consistently rank in the Top 10, his AI and Law writings are globally recognized.

References

[Basitat, 1850] Frederic Basitat. *The Law.* Ludwig von Mises Institute, Auburn, Alabama, 1850.

[Bench-Capon, 2020] T.J.M. Bench-Capon. Ethical approaches and autonomous systems. *Artificial Intelligence*, 281(3): 1-15, January 2020.

[Bickenbach, 1989] Jerome Bickenbach. Law and morality. *Law and Philosophy,* 3(3): 291-300, March 1989.

[Brandao et al, 2020] Martim Brandao, Marina Jirotka, Helena Webb, and Paul Luff. Fair navigation planning: A resource for characterizing and designing fairness in mobile robots. *Artificial Intelligence*, 282(1): 1-20, March 2020.

[Cane, 2012] Peter Cane. Morality, law and conflicting reasons for action. *The Cambridge Law Journal*, 71(1): 59-85, March 2012.

[Dicey, 1885] Albert Dicey. *Introduction to the Study of the Law of the Constitution*. McMillan and Co., London, 1885.

[Eliot, 2016] Lance Eliot. *AI Guardian Angels For Deep AI Trustworthiness*. LBE Press Publishing, Los Angeles, California 2016.

[Eliot, 2017] Lance Eliot. Self-Aware AI. *AI Trends*, 2(8): 1-5, August 2017.

[Eliot, 2018] Lance Eliot. Probabilistic reasoning and AI. *AI Trends*, 3(2): 14-17, February 2018.

[Eliot, 2018] Lance Eliot. AI arguing machines. *AI Trends*, 3(11): 1-8, November 2018.

[Eliot, 2019] Lance Eliot. *AI and Legal Reasoning Essentials*. LBE Press Publishing, Los Angeles, California 2019.

[Freedman et al, 2020] Rachel Freedman, Jana Borg, Walter Sinnott-Armstrong, John Dickerson, and Vincent Conitzer. Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence*, 283(2): 1-14, March 2020.

[Fridman, 2017]. Lex Fridman, Li Ding, Benedikt Jenik, and Bryan Reimer. Arguing machines. *Cornell University AI arXiv*, 1710.04459, October 2017.

[Genesereth, 2009] Michael Genesereth. Computational law. *Stanford Center for Legal Informatics*, May 2019.

[Hage et al, 1993] Jaap Hage, Ronald Leenes, and Arno Lodder. Hard cases: A procedural approach. *Artificial Intelligence and Law*, 82(1): 52-67, February 1993.

[Huang, 2019] Bert Huang. Law's halo and the moral machine. *Columbia Law Review,* 119(7): 1811-1828, November 2019.

[Kagan, 1998] Shelly Kagan. *The Limits of Morality*. Oxford Scholarship Online, November 1998.

[Maslow, 1943] Abraham Maslow. A theory of human motivation. *Psychological Review*, 50(4): 370-396, April 1943.

[Misselhorn, 2019] Catrin Misselhorn. Artificial systems with moral capacities. *Artificial Intelligence*, 278(1): 1-11, October 2019.

[Mnookin and Kornhauser, 1979] Robert Mnookin and Lewis Kornhauser. Bargaining in the shadow of law. *The Yale Law Review,* 88(5): 950-997, April 1979.

[Parasuraman et al, 2000] Raja Parasuraman, Thomas Sheridan, and Christopher Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics*, 30(3): 286-297, May 2000.

[Rao, 1980] Goparaju Ramachandra Rao. *I Learn.* Navjivan Trust, Ahmedabad, India, 1980.

[Regan, 1987] Donald Regan. Law's halo. *Social Philosophy and Policy*, 4(1): 15-30, April 1987.

[Shiell, 1987] Timothy Shiell. Making sense out of a necessary connection between law and morality. *Public Affairs Quarterly*, 1(3): 77-90, July 1987.

[Surden, 2019] Artificial Intelligence and the law: An overview. *Georgia State University Law Review*, 8(2): 41-57, August 2019.

[Waltl and Vogl, 2018] Bernhard Waltl and Roland Vogl. Increasing transparency in algorithmic decision-making with explainable AI. *Datenschutz Datensich*, 42(1): 613-617, October 2018.