

## MIT Open Access Articles

*Society-in-the-loop: programming the algorithmic social contract*

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

**Citation:** Rahwan, Iyad. "Society-in-the-Loop: Programming the Algorithmic Social Contract." Ethics and Information Technology 20, no. 1 (August 17, 2017): 5–14.

**As Published:** <http://dx.doi.org/10.1007/s10676-017-9430-8>

**Publisher:** Springer Netherlands

**Persistent URL:** <http://hdl.handle.net/1721.1/114718>

**Version:** Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

**Terms of use:** Creative Commons Attribution-Noncommercial-Share Alike



# Society-in-the-Loop: Programming the Algorithmic Social Contract

Iyad Rahwan<sup>1,2,\*</sup>

<sup>1</sup>The Media Lab, Massachusetts Institute of Technology, Cambridge MA, 02139 USA

<sup>2</sup>Institute for Data, Systems and Society, Massachusetts Institute of Technology, Cambridge MA, 02139 USA

This version: 20 July 2017

**Abstract** Recent rapid advances in Artificial Intelligence (AI) and Machine Learning have raised many questions about the regulatory and governance mechanisms for autonomous machines. Many commentators, scholars, and policy-makers now call for ensuring that algorithms governing our lives are transparent, fair, and accountable. Here, I propose a conceptual framework for the regulation of AI and algorithmic systems. I argue that we need tools to program, debug and maintain an *algorithmic social contract*, a pact between various human stakeholders, mediated by machines. To achieve this, we can adapt the concept of *human-in-the-loop* (HITL) from the fields of modeling and simulation, and interactive machine learning. In particular, I propose an agenda I call *society-in-the-loop* (SITL), which combines the HITL control paradigm with mechanisms for negotiating the values of various stakeholders affected by AI systems, and monitoring compliance with the agreement. In short, ‘*SITL = HITL + Social Contract*.’

## 1 Introduction

*“Art goes yet further, imitating that Rationall and most excellent worke of Nature, Man. For by Art is created that great LEVIATHAN called a COMMON-WEALTH, or STATE, (in latine CIVITAS) which is but an Artificiall Man”*

Thomas Hobbes (1651). Leviathan

Despite the initial promise of Artificial Intelligence, a long ‘AI Winter’ ensued in the 1980s and 1990s, as problems of automated reasoning proved much harder

than initially anticipated [48]. But recent years have seen rapid theoretical and practical advances in many areas of AI. Prominent examples include machines learning their own representations of the world via Deep Neural Network architectures [41], Reinforcement Learning from evaluative feedback [45], and economic reasoning in markets and other multi-agent systems [58]. The result is an accelerating proliferation of AI technologies in everyday life [43].

These advances are yielding substantial societal benefits, ranging from more efficient supply chain management, to better matchmaking in peer-to-peer markets and online dating apps, to more reliable medical diagnosis and drug discovery [71].

But AI advances have also raised many questions about the regulatory and governance mechanisms for autonomous machines and complex algorithmic systems. Some commentators are concerned that algorithmic systems are not accountable because they are *black boxes* whose inner workings are not transparent to all stakeholders [59]. Others raised concern over people unwittingly living in filter bubbles created by news recommendation algorithms [11, 57]. Others argue that data-driven decision-support systems can perpetuate injustice, because they can also be biased either in their design, or by picking up human biases in their training data [13, 76]. Furthermore, algorithms can create feedback loops that reinforce inequality [10], for example in the use of AI in *predictive policing* or *creditworthiness* prediction, making it difficult for individuals to escape the vicious cycle of poverty [54].

In response to these alarms, various academic and governmental entities have started thinking seriously about AI governance. Recently, the United States White House National Science and Technology Council Committee on Technology released a report with recom-

\*An earlier version of this article was published under the same title on [medium.com](https://medium.com), on August 12, 2016.

<sup>a</sup>e-mail: irahwan@mit.edu

recommendations ranging from eliminating bias from data, to regulating autonomous vehicles, to introducing ethical training to computer science curricula [51]. The European Union, which has enacted many personal data privacy regulations, will soon vote on a proposal to grant robots legal status in order to hold them accountable, and to produce a code of ethical conduct for their design [21]. The Institute of Electrical and Electronics Engineers recently published a vision on ‘Ethically Aligned Design’ [37]. Industry leaders have also taken the initiative to create a ‘Partnership on AI’ to establish best practices for AI systems and to educate the public about AI [34].

My goal in this paper is to introduce a conceptual framework for thinking about the regulation of AI and data-driven systems. I argue that we need a new kind of social contract: an *algorithmic social contract*, that is a contract between various stakeholders, mediated by machines. To achieve this, we need to adopt a *society-in-the-loop* (SITL) framework in thinking about AI systems, which adapts the concept of *human-in-the-loop* (HITL), from the fields of supervisory control and interactive machine learning, but extends it to oversight conducted by society as a whole.

## 2 Human-in-the-Loop

In a *human-in-the-loop* (HITL) system, a human operator is a crucial component of an automated control process, handling challenging tasks of supervision, exception control, optimization and maintenance (Figure 1). The notion has been studied for decades within the field of *supervisory control* [2, 68]. Sheridan defined *human supervisory control* as a process by which “one or more human operators are intermittently programming and continually receiving information from a computer that itself closes an autonomous control loop through artificial effectors to the controlled process or task environment” [67].

These ideas then made their way into the field of Human-Computer Interaction (HCI). Scientists began working on *mixed-initiative* user interfaces, in which the autonomous system can make intelligent decisions about when and how to engage the human [36].

Recently, a number of articles have been written about the importance of applying HITL thinking to Artificial Intelligence (AI) and machine learning (ML) systems. A simple form of HITL ML is the use of human workers to label data for training machine learning algorithms. This has produced invaluable benchmarks that spurred major advances in computer vision, for example [65].

Another example of HITL ML is *interactive machine learning*, which can help machines learn faster or more effectively by integrating feedback interactively from users [3, 20]. This type of HITL ML has been going on for a while. For example, many computer applications learn from your behavior in order to improve their ability to serve you better (e.g. by predicting the next word you are going to type). Similarly, when you mark an email as ‘spam’ in an online email service, you are one of many humans in the loop of a complex machine learning algorithm (specifically an active learning system), helping it in its continuous quest to improve email classification as spam or non-spam.

More sophisticated examples of HITL ML are now emerging, in which the human-in-the-loop has more explicit knowledge of the state of the system. For instance, in a crisis counseling system, a machine learning system classifies messages sent by callers, and provides visualizations to a human counselor in real-time [23]. Thus, the human and the machine learning system work in tandem to deliver effective counseling.

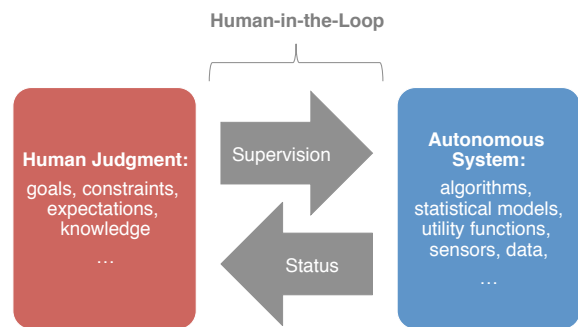


Fig. 1: In a HITL system, a human provides monitoring and supervisory functions at crucial junctions in the system’s operation.

HITL thinking has also been applied successfully to human-robot interaction (HRI) [12]. This includes dynamically adapting the degree of autonomy given to robots [19, 73], interactively teaching reinforcement learning robots to adopt particular behaviors [74], and designing flexible human-robot teams [38].

There is another role of the HITL paradigm, which is closer to the problems discussed in the present article. HITL is not only a means to *improve* AI systems’ accuracy in classification or to speed up the convergence of a reinforcement learning robot. Rather, HITL can also be a powerful tool for *regulating* the behavior of AI systems. For instance, many scholars now advocate for expert oversight, by a human operator, over the behavior of ‘killer robots’ or credit scoring algorithms [17].

The presence of a human fulfills two major functions in a HITL AI system:

1. The human can *identify misbehavior* by an otherwise autonomous system, and take corrective action. For instance, a credit scoring system may misclassify an adult as ineligible for credit, due to an error in data entry in their age—something a human may spot from the applicant’s photograph. Similarly, a computer vision system on a weaponized drone may mis-identify a civilian as a combatant, and the human operator—it is hoped—would ensure that such cases are identified, and override the system. Some work is underway to ensure AI cannot learn to disable their own kill-switch [56].
2. The human can be involved in order to provide an *accountable entity* in case the system misbehaves. If a fully autonomous system causes harm to human beings, having a human in the loop provides trust that somebody would bare the consequence of such mistakes, and thus have incentive to minimize those mistakes. This person may be a human within a tight control loop (e.g. an operator of a drone) or a much slower loop (e.g. programmers in a multi-year development cycle of an autonomous vehicle). Until we find a way to punish algorithms for harm to humans, it is hard to think of any other alternative.

While HITL is a useful interaction paradigm for building AI systems that are subject to oversight, I believe it does not sufficiently emphasize the role of society as a whole in such oversight. HITL suggests that once we put a human expert, or group of experts, within the loop of an AI system, the problem of regulation is solved. But as I shall discuss in the following section, this may not always be the case.

### 3 Society-in-the-Loop

What happens when an AI system does not serve a narrow, well-defined function, but a broad function with wide societal implications? Consider an AI algorithm that controls millions of self-driving cars; or a set of news filtering algorithms that influence the political beliefs and preferences of millions of citizens; or algorithms that mediate the allocation of resources and labor in an entire economy. What is the HITL equivalent of these algorithms? This is where we make the qualitative shift from HITL to *society in the loop* (SITL).

While HITL AI is about embedding the judgment of *individual* humans or groups in the optimization of AI systems with *narrow impact*, SITL is about embedding the values of *society*, as a whole, in the algorithmic

governance of societal outcomes that have *broad implications*. In other words, SITL becomes relevant when the scope of both the input and the output of AI systems is very broad. But one might ask, why should this be any different?

The move from HITL to SITL raises a fundamentally different problem: how to balance the competing interests of different stakeholders, including the interests of those who govern through algorithms? This is, traditionally, a problem of defining a *social contract* [70]. To put it in the most skeletal form, we can say:

$$\text{SITL} = \text{HITL} + \text{Social Contract}$$

To elaborate on this simple equation, we need to take a short detour into political philosophy.

### 4 Detour: The Social Contract

Humans are the ultimate cooperative species [53]. Cultural anthropologists trace the evolution of political systems of governance from decentralized bands and tribes, to increasingly centralized chiefdoms, sovereign states and empires [31].

Over time, humans reached the limits of old cooperative institutions such as *kin selection*—helping others who share their genes [30], and *reciprocal altruism*—helping others who would later help them back [75]. These old mechanisms cannot scale adequately to larger groups. In the face of inter-group competition, evolutionary pressure favored the emergence, and spread, of more complex social institutions to coordinate people’s behaviors [77, 80]. For example, centralized sanctioning power is able to prevent higher-order free-riding—following cooperative norms, but not contributing to their enforcement—that undermine cooperation in larger groups [6, 29, 69].

The founders of *social contract* theory, going back to Thomas Hobbes’ landmark book, *Leviathan* [35], posit that centralized government is legitimate precisely because it enables industrious people to cooperate via third-party enforcement of contracts among strangers (see Figure 2). Some of these contracts are explicit, such as marriage contracts or commercial transactions. Other aspects of the social contract are implicit, being embedded in social norms that govern every day life. In both cases, the contract embodies mutual consent to the government’s legitimate use of force—or people’s use of social pressure—to guard people’s rights and punish violators [8, 70].

Hobbes gave his *Leviathan*, the sovereign, enormous power. Subsequently, the social contract undertook many stages of evolution, thanks to enlightenment thinkers like John Locke [47], Jean-Jacques Rousseau





Fig. 2: The frontispiece of Thomas Hobbes’ 1651 book *Leviathan* by Parisian artist Abraham Bosse. The piece is a striking depiction of how the sovereign—a giant ruling a peaceful realm through the warrior’s sword and the monk’s crosier. The torso and arms of the figure are composed of over three hundred persons, signifying that the Leviathan derives his power to govern, not from divine authority, but through the consent of the governed.

[64], all the way to John Rawls [62] David Gauthier [28] and Brian Skyrms [70] in modern times. These thinkers refined our conception of how the social contract emerges in the first place, as well as the ways in which we can keep it from collapsing.

Modern political institutions, including the modern state, are a product of these evolutionary mechanisms of political development, which combine institutional innovation with learning. As Fukuyama puts it, “[s]ocieties are not trapped by their pasts and freely borrow ideas and institutions from each other” [26].

The result of this evolutionary process is a social contract that can provide the efficiency and stability of sovereign states, but which also ensures the sovereign implements the *general will* [64] of the people, and is held in some way accountable for violations of fundamental rights. In the same manner, “[n]ew algorithmic decisionmakers are sovereign over important aspects of individual lives” Thus, lack of accountability and due process for algorithmic decisions risks “paving the way to a new feudal order” [17].

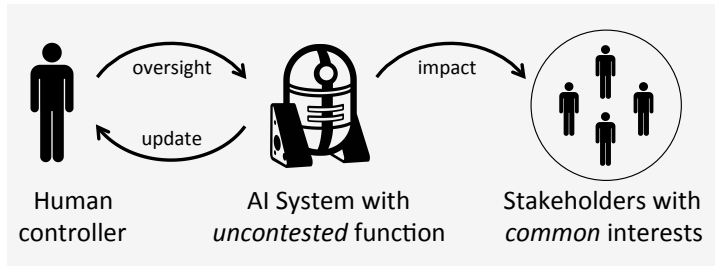
## 5 The Algorithmic Social Contract

The SITL paradigm that I advocate is more akin to the interaction between a government and a governed citizenry, than the interaction between a drone and its operator. Similar to the role of due process and accountability in the traditional social contract, SITL can be conceived as an attempt to embed the general will into an *algorithmic* social contract.

By using the social contract metaphor, the SITL paradigm emphasizes an important distinction from the traditional HITL paradigm (Figure 3). In a HITL system, a human controller ensures that the AI system fulfills *uncontested* and *common* goals on behalf of societal stakeholders—e.g. ensuring a plane lands safely, or improving food quality inspection. In addition, in the SITL domain, society must agree on two aspects:

1. Society must resolve tradeoffs between the different values that AI systems can strive towards—e.g. tradeoffs between security and privacy, or the tradeoffs between different notions of fairness [7, 39].

### Human-in-the-Loop (HITL)



### Society-in-the-Loop (SITL)

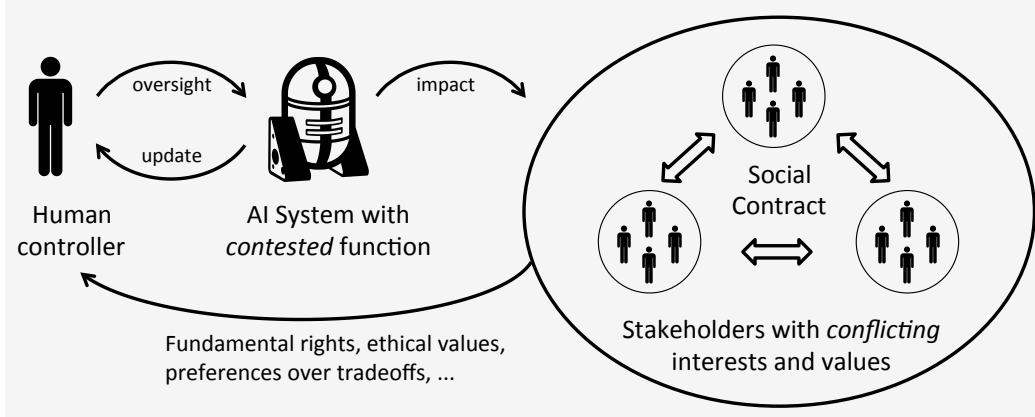


Fig. 3: Society-in-the-Loop (SITL) = Human-in-the-Loop (HITL) + Social Contract; (*Top*) In a HITL system, a human controller monitors and exercises oversight over the operation of an AI system to ensure that it serves the uncontested and common goals of its stakeholders. For example, a human pilot oversees an airplane autopilot to increase passenger safety. (*Bottom*) In SITL, the AI system has broad impact, requiring various societal stakeholders to identify the fundamental rights that the AI must respect, the ethical values that should guide the AI's operation, the cost and benefit tradeoffs the AI can make between various stakeholder groups, etc.

2. Society must agree on which stakeholders would reap which benefits and pay which costs—e.g. how improvements in safety made possible by driverless cars are to be distributed between passengers and pedestrians, or which degree of collateral damage, if any, is acceptable in autonomous warfare.

In human-based government, citizens use various channels—e.g. democratic voting, opinion polls, civil society institutions, the media—to articulate their expectations to the government. Meanwhile, the government, through its bureaucracy and various branches undertakes the function of governing, and is ultimately evaluated by the citizenry. And while citizens are not involved in the details [44], they are the arbiters among all of these institutions, and have the power to replace their key actors.

Modern societies are (in theory) SITL human-based governance machines. Some of those machines are better programmed, and have better ‘user interfaces’ than others. Similarly, as more governance functions get en-

coded into AI algorithms, we need to create channels between human values and governance algorithms.

To implement SITL, we need to know what types of behaviors people expect from AI, and to enable policy-makers and the public to articulate these expectations (goals, ethics, norms, social contract) to machines. To close the loop, we also need new metrics and methods to evaluate AI behavior against quantifiable human values. In other words: we need to build new tools to program, debug, and monitor the algorithmic social contract between humans and algorithms—that is, algorithms that are effective sovereigns over important aspects of social and economic life, whether or not they are actually operated by governments. This requires both government regulation and industry standards that represent the expectations of the public, with corresponding oversight.

## 6 The SITL Gap

Why are we not there yet? There has been a flurry of thoughtful treaties on the social and legal challenges posed by the opaque algorithms that permeate and govern our lives. While these seminal writings help illuminate many of the challenges, they fall short on comprehensive solutions.

### 6.1 Articulating Societal Values

One barrier to implementing SITL is the cultural divide between engineering on one hand, and the humanities on the other (see Figure 4). Thoughtful legislators, legal scholars, media theorists, and ethicists are very skilled at revealing moral hazards, and identifying ways in which moral principles and constitutional rights may be violated [15]. But they are not always able to articulate those expectations in ways that engineers and designers can operationalize.

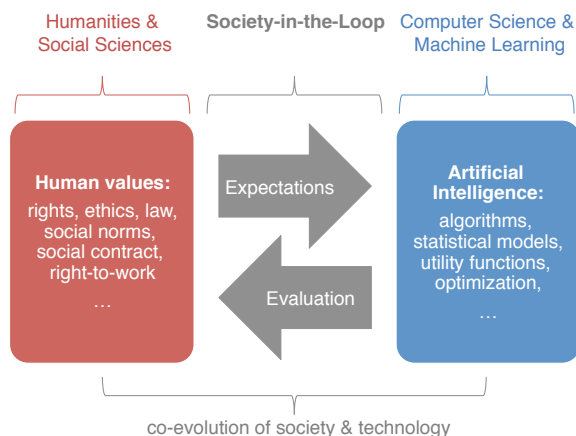


Fig. 4: In a SITL system, broad societal values (as opposed to an individual human operator’s judgment) must be involved in the monitoring and supervisory function of AI systems that have wide-ranging societal implications (as opposed to AI systems with narrow impact).

### 6.2 Quantifying Externalities & Negotiating Tradeoffs

Algorithms can generate what economists refer to as *negative externalities*—costs incurred by third parties not involved in the decision [61]. For example, if autonomous vehicle algorithms over-prioritize the safety of passengers—who own them or pay to use them—they may disproportionately increase the risk borne by

pedestrians. Quantifying these kinds of externalities is not always straightforward, especially when they occur as a consequence of long, indirect causal chains, or as a result of machine code that is opaque to humans.

Once we have quantified externalities, we need to negotiate the *tradeoffs* they embody. If certain ways to increase pedestrian safety in autonomous vehicles imply reduction in passenger safety, which tradeoffs are acceptable?

Human experts already implement tradeoffs as they design policies and products. For example, reducing the speed limit on a road reduces the utility for drivers who want to get home quickly, while increasing the overall safety of drivers and pedestrians. It is possible to completely eliminate accidents—by reducing the speed limit to zero and banning cars—but this would also eliminate the utility of driving, and regulators attempt to strike a balance that society is comfortable with through a constant learning process.

Quantifying tradeoffs in any complex system, with many interacting parts, is always difficult. In complex economic systems, there are often *unintended consequences* of design choices. As AI becomes an integral part of such systems, the problem of quantifying those tradeoffs becomes even harder. For example, subtle algorithm design choices in autonomous vehicles may lead to a particular tradeoff between risks to passengers and risks to pedestrians. Identifying, let alone negotiating those tradeoffs, may be much harder than setting a speed limit—if only due to the greater degrees of freedom when making design choices. This may be further complicated by the fact that algorithms learn from their experience, which may lead to shifts in the tradeoffs being made, going beyond what the programmers intended.

### 6.3 Verifying Compliance with Societal Values

Computer scientists and engineers are not always able to quantify the behaviors of their systems such that they can be easily understood by ethicists and legal theorists. This makes it more difficult to scrutinize the behavior of algorithms against set expectations. Even simple notions such as ‘fairness’ can be formalized in many different ways mathematically or in computer code [7].

An important component of Figure 4 is that both human values and AI are ongoing constant co-evolution. Thus, the evolution of technical capability can dramatically (and even irreversibly) alter what society considers acceptable—think of how privacy norms have changed because of the utility provided by smart phones and the Internet.



## 7 Bridging the Gap

There are many efforts underway to bridge the society-in-the-loop gap. Below is an incomplete list of efforts that I believe are relevant, and a discussion of their merits and limitations.

### 7.1 Articulating Values: Design, Crowdsourcing & Sentiment Analysis

In the broader context of technology design, various *value-sensitive design* methodologies have been proposed [25], which can be applied to software development [1, 79]. These approaches may prove helpful in the design of AI systems.

Some AI scientists propose to use of crowdsourcing [18] to identify societal tradeoffs in a programmable way. There are some efforts to collect data about people's preferences over values implemented in AI algorithms, such as those that control driverless cars. Using methods from the field of moral psychology, one can identify potential moral hazards due to the incentives of different users of the road [9]. For example, my co-authors and I have developed a public-facing survey tool that elicits the public's moral expectations from autonomous cars faced with ethical dilemmas [49]. We have collected over 30 million decisions to date. Findings from this data can help regulators and car makers understand some of the psychological barriers to the wide adoption of autonomous vehicles.

In many domains, it may be possible to measure societal values directly from observational data, without having to run explicit polling campaigns or build dedicated crowdsourcing platforms [46]. For example, automated sentiment analysis on social media discourse can quantify people's reaction to different moral violations committed by AI systems. While these approaches have their limitations, they can help gauge the evolution of public attitudes, and their readiness to accept new social pacts through machines.

### 7.2 Negotiation: Social Choice & Contractarianism

The field of *computational social choice* [4, 50] explores the aggregation of societal preferences and fair allocation of resources. Because these aggregation mechanisms can be implemented algorithmically, they provide a potential solution to the problem of negotiating tradeoffs of different stakeholders [16, 52, 58].

An alternative approach to the negotiation of values is to use normative and meta-ethical tools from social contract theory to identify enforceable outcomes

that rational actors would be willing to opt into. For instance, Leben recently proposed an algorithm that allows autonomous vehicles to resolve dilemmas of unavoidable harm using Rawls' Contractarianism [40]. In particular, Leben proposes to program cars to make decisions that rational actors would take if they were in a hypothetical 'original position' behind a 'veil of ignorance.' This veil would, for example, conceal whether the person is a passenger or a pedestrian in a given accident, leading them to choose the *maximin* solution—that is, the decision that minimizes how bad the worse-case outcome is.

### 7.3 Compliance: People Watching Algorithms

An important function for ensuring accountability is the ability to scrutinize the behavior of those in power, through mechanisms of transparency. In the context of algorithms, this does *not* mean having access to computer source code, as intuitive as this notion might seem.

Reading the source code of a modern machine learning algorithm tells us little about its behavior, because it is often through the interaction between algorithms and data that things like discrimination emerge. Transparency must, therefore, be about the external behavior of algorithms. Indeed, this is how we regulate the behavior of humans—not by looking into their brain's neural circuitry, but by observing their behavior and judging it against certain standards of conduct. Of course, this observation can benefit from the ability of the algorithm to give human-interpretable explanations of their decisions [42].

The new journalistic practice of *algorithmic accountability reporting* provides a framework for scrutiny of algorithmic decisions that is purely behavioral [22]. As an example, Sweeney has demonstrated that Web searches for names common among African Americans cause online advertising algorithms to serve ads suggestive of an arrest record, which can harm the individual being searched [72]. Investigative journalism has also revealed evidence of price discrimination based on users' information, sparking a debate about the appropriateness of this practice [78].

We might also envision a role for professional *algorithm auditors*, people who interrogate algorithms to ensure compliance with pre-set standards. This interrogation may utilize real or synthetic datasets designed to identify whether an algorithm violates certain requirements. For instance, an algorithm auditor may provide sample job applications to identify if a job matching algorithm is discriminating between candidates based on irrelevant factors. Or an autonomous vehicle algorithm auditor may provide simulated traffic scenarios to en-



sure the vehicle is not disproportionately increasing the risk to pedestrians or cyclists in favor of passengers.

One weakness of auditing in a simulated environment—using computer simulation or fake data—is the potential for adversarial behavior: the algorithm being audited may attempt to trick the algorithm doing the auditing. This is similar to ‘defeat devices,’ a term used to describe software or hardware features that interfere with or disables car emissions controls under real world driving conditions, even if the vehicle passes formal emissions testing [27]. In a similar fashion, an autonomous vehicle control algorithm may detect that it is being tested in a virtual environment—e.g. by noticing that the distribution of scenarios is skewed towards ethical dilemmas—and behave differently under such testing conditions.

The possibility of this generalized ‘defeat device’ subversion necessitates continuous monitoring and auditing in real-world conditions, not just simulated conditions at certification time. Such continuous monitoring may benefit from automation, as I discuss in the next section.

#### 7.4 Compliance: Algorithms Watching Algorithms

Recently, Amitai and Oren Etzioni proposed a new class of algorithms, called *oversight programs*, whose function is to “monitor, audit, and hold operational AI programs accountable” [24]. Note the emphasis on ‘operational,’ suggesting that these oversight programs are aligned with the point I made earlier about the futility of source code inspection as the only means for regulation.

Oversight algorithms, thus, perform a similar function to today’s spam filtering algorithms. But their scope is much wider, as they investigate suspicious behavior by rogue AI algorithms maliciously violating human values. For example, a new class of browser plug-ins is allowing independent, data-driven auditing of the information provided by online advertising platforms to advertisers [14]. This has revealed issues in the transparency and accuracy of the current algorithmically-mediated online advertising ecosystem.

One can imagine an algorithm that conducts real-time quantification of the amount of bias caused by a news filtering algorithm—akin to Facebook’s recent study [5]—and raising an alarm if bias increases beyond a certain threshold.

#### 7.5 The Limits of Public Engagement

It is worth highlighting the limits of crowdsourcing of societal values in general, and when it comes to

AI in particular. One of the most influential figures in 20th century journalism, Walter Lippman, warned of over-reliance on public opinion when it comes to policy matters that require significant expertise. In Lippman’s words, “Public opinion is not a rational force.... It does not reason, investigate, invent, persuade, bargain or settle” [44]. This is because it is impossible for a lay person to be fully informed about all facets of every policy question: even an expert practitioner or regulator in one field—say medicine—cannot be sufficiently informed to weigh in on policy matters in another field—say monetary policy. The role of public opinion, Lippman contends, is to check the use of sovereign force, based on assessments made digestible to them by disagreeing experts, pundits and journalists.

There is a lot of merit in Lippman’s argument. But he misses a second important role that the public plays: that of shaping moral values and norms. Experts alone cannot dictate what societal values should be. They can influence those values by providing relevant facts, such as the importance of physical exercise in promoting health, or the importance of recycling in the preservation of the environment. But ultimately, norms are shaped through the interaction of various social and evolutionary forces [33, 63]. And these values must influence the metrics against which the performance of experts—or AI algorithms—is measured.

### 8 Discussion

The ideas outlined in this article are not entirely new, and many have been discussed in the context of digital democracy [32] and the data-driven society [60]. Tim O’Reilly recently coined the term *algorithmic regulation* to describe data-driven governance [55]. To O’Reilly, successful algorithmic regulation must satisfy the following properties (quoted verbatim):

1. A deep understanding of the desired outcome
2. Real-time measurement to determine if that outcome is being achieved
3. Algorithms (i.e. a set of rules) that make adjustments based on new data
4. Periodic, deeper analysis of whether the algorithms themselves are correct and performing as expected.

I agree with O’Reilly’s characterization. From my perspective, the identification and negotiation of desired outcomes are non-trivial problems. And ensuring that algorithms are performing as expected is not just a technical challenge, but also a social one. This is what makes the social contract framework helpful.

Note that SITL operates at different time-scales than HITL. It looks more like public feedback on regula-

tions and legislations, than feedback on frequent micro-level decisions. Nevertheless, I believe there is value in ensuring we pay attention to all component of ‘the loop’ using an explicit framework. This will be increasingly important as the time between diagnosis and policy adjustment becomes shorter, thanks to progress in data science and machine learning.

I attempted to synthesize various concerns and solutions put forward by many scholars who are thinking about the regulation of algorithmic systems that govern social and economic life. I organized these discussions within two paradigms that have a long history: the human-in-the-loop paradigm from the fields of computer science and supervisory control, and the ‘social contract’ paradigm from political philosophy. The result can be summarized by a call-to-arms that defines the challenge ahead:

*to build institutions and tools that put the society-in-the-loop of algorithmic systems, and allows us to program, debug, and monitor the algorithmic social contract between humans and governance algorithms.*

The Age of Enlightenment marked humanity’s transition towards the modern social contract, in which political legitimacy no longer emanates from the divine authority of kings, but from the mutual agreement among free citizens to appoint a sovereign. We spent centuries taming Hobbes’s Leviathan, the all-powerful sovereign [35]. We must now create and tame the new *Techno-Leviathan*.

## Acknowledgement

I am grateful for financial support from the Ethics & Governance of Artificial Intelligence Fund, as well as support from the Siegel Family Endowment.

I am indebted to Joi Ito, Suelette Dreyfus, Cesar Hidalgo, Alex ‘Sandy’ Pentland, Tenzin Priyadarshi and Mark Staples for conversations and comments that helped shape this article. I’m grateful to Brett Scott for allowing me to appropriate the term ‘Techno-Leviathan’ which he originally presented in the context of Cryptocurrency [66]. I thank Deb Roy for introducing me to Walter Lippman’s ‘The Phantom Public’ and for constantly challenging my thinking. I thank Danny Hillis for pointing to the co-evolution of technology and societal values. I thank James Guszcza for suggesting the term ‘algorithm auditors’ and for other helpful comments.

## References

1. Aldewereld, H., Dignum, V., and hua Tan, Y. (2014). Design for values in software development. In Jeroen van den Hoven, Pieter E. Vermaas, I. v. d. P., editor, *Handbook of Ethics, Values, and Technological Design*. Springer.
2. Allen, J., Guinn, C. I., and Horvitz, E. (1999). Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications*, 14(5):14–23.
3. Amershi, S., Cakmak, M., Knox, W. B., and Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120.
4. Arrow, K. J. (2012). *Social choice and individual values*, volume 12. Yale university press.
5. Bakshy, E., Messing, S., and Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132.
6. Baldassarri, D. and Grossman, G. (2011). Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Sciences*, 108(27):11023–11027.
7. Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2017). Fairness in criminal justice risk assessments: The state of the art. *arXiv preprint arXiv:1703.09207*.
8. Binmore, K. (2005). *Natural justice*. Oxford University Press.
9. Bonnefon, J.-F., Shariff, A., and Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, 352(6293):1573–1576.
10. Boyd, D. and Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679.
11. Bozdog, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and information technology*, 15(3):209–227.
12. Cakmak, M., Chao, C., and Thomaz, A. L. (2010). Designing interactions for robot active learners. *IEEE Transactions on Autonomous Mental Development*, 2(2):108–118.
13. Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.
14. Callejo, P., Cuevas, R., Cuevas, A., and Kotila, M. (2016). Independent auditing of online display advertising campaigns. In *Proceedings of the 15th ACM Workshop on Hot Topics in Networks (HotNets)*, pages 120–126.
15. Castelfranchi, C. (2000). Artificial liars: Why computers will (necessarily) deceive us and each other. *Ethics and Information Technology*, 2(2):113–119.
16. Chen, Y., Lai, J. K., Parkes, D. C., and Procaccia, A. D. (2013). Truth, justice, and cake cutting. *Games and Economic Behavior*, 77(1):284–297.
17. Citron, D. K. and Pasquale, F. A. (2014). The scored society: due process for automated predictions. *Washington Law Review*, 89.
18. Conitzer, V., Brill, M., and Freeman, R. (2015). Crowdsourcing societal tradeoffs. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1213–1217. International Foundation for Autonomous Agents and Multiagent Systems.
19. Crandall, J. W. and Goodrich, M. A. (2001). Experiments in adjustable autonomy. In *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*, volume 3, pages 1624–1629. IEEE.
20. Cuzzillo, T. (2015). Real-world active learning: Applications and strategies for human-in-the-loop machine learning. Technical report, OaŽReilly.

21. Delvaux, M. (2016). Motion for a European Parliament resolution: with recommendations to the commission on civil law rules on robotics. Technical Report (2015/2103(INL)), European Commission.
22. Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism*, 3(3):398–415.
23. Dinakar, K., Chen, J., Lieberman, H., Picard, R., and Filbin, R. (2015). Mixed-initiative real-time topic modeling & visualization for crisis counseling. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 417–426. ACM.
24. Etzioni, A. and Etzioni, O. (2016). AI assisted ethics. *Ethics and Information Technology*, 18(2):149–156.
25. Friedman, B. (1996). Value-sensitive design. *interactions*, 3(6):16–23.
26. Fukuyama, F. (2011). *The origins of political order: from prehuman times to the French Revolution*. Profile books.
27. Gates, G., Ewing, J., Russell, K., and Watkins, D. (2015). How Volkswagen’s ‘defeat devices’ worked. *New York Times*.
28. Gauthier, D. (1986). *Morals by agreement*. Oxford University Press on Demand.
29. Gülerk, Ö., Irlenbusch, B., and Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science*, 312(5770):108–111.
30. Hamilton, W. D. (1963). The evolution of altruistic behavior. *American naturalist*, pages 354–356.
31. Haviland, W., Prins, H., McBride, B., and Walrath, D. (2013). *Cultural anthropology: the human challenge*. Cengage Learning.
32. Helbing, D. and Pournaras, E. (2015). Society: Build digital democracy. *Nature*, 527:33–34.
33. Henrich, J. (2004). Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior & Organization*, 53(1):3–35.
34. Hern, A. (2016). ‘partnership on artificial intelligence’ formed by Google, Facebook, Amazon, IBM, Microsoft and Apple. Technical report, The Guardian.
35. Hobbes, T. (1651). *Leviathan, or, the Matter, Forme, and Power of a Common-Wealth Ecclesiasticall and Civill*.
36. Horvitz, E. (1999). Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 159–166. ACM.
37. IEEE (2016). Ethically aligned design. Technical report, The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems.
38. Johnson, M., Bradshaw, J. M., Feltoovich, P. J., Jonker, C. M., Van Riemsdijk, M. B., and Sierhuis, M. (2014). Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction*, 3(1):43–69.
39. Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.
40. Leben, D. (2017). A rawlsian algorithm for autonomous vehicles. *Ethics and Information Technology*, 19:107–115.
41. LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.
42. Letham, B., Rudin, C., McCormick, T. H., Madigan, D., et al. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371.
43. Levy, S. (2010). The AI revolution is on. *Wired*.
44. Lippmann, W. (1927). *The phantom public*. Transaction Publishers.
45. Littman, M. L. (2015). Reinforcement learning improves behaviour from evaluative feedback. *Nature*, 521(7553):445–451.
46. Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.
47. Locke, J. (1689). *Two treatises of government*. Self Published.
48. Markoff, J. (2015). *Machines of loving grace*. Ecco.
49. MIT (2017). The moral machine. <http://moralmachine.mit.edu>. Accessed: 2017-01-01.
50. Moulin, H., Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D. (2016). *Handbook of Computational Social Choice*. Cambridge University Press.
51. National Science and Technology Council Committee on Technology (2016). Preparing for the future of artificial intelligence. Technical report, Executive Office of the President.
52. Nisan, N., Roughgarden, T., Tardos, E., and Vazirani, V. V. (2007). *Algorithmic game theory*, volume 1. Cambridge University Press Cambridge.
53. Nowak, M. and Highfield, R. (2011). *Supercooperators: Altruism, evolution, and why we need each other to succeed*. Simon and Schuster.
54. O’Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishing Group (NY).
55. O’Reilly, T. (2016). Open data and algorithmic regulation. In Goldstein, B. and Dyson, L., editors, *Beyond Transparency: Open Data and the Future of Civic Innovation*. Code for America Press.
56. Orseau, L. and Armstrong, S. (2016). Safely interruptible agents. In *Uncertainty in Artificial Intelligence: 32nd Conference (UAI)*.
57. Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin UK.
58. Parkes, D. C. and Wellman, M. P. (2015). Economic reasoning and artificial intelligence. *Science*, 349(6245):267–272.
59. Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.
60. Pentland, A. S. (2013). The data-driven society. *Scientific American*, 309(4):78–83.
61. Pigou, A. C. (1920). *The economics of welfare*. Palgrave Macmillan.
62. Rawls, J. (1971). *A theory of justice*. Harvard university press.
63. Richerson, P. J. and Boyd, R. (2005). Not by genes alone.
64. Rousseau, J.-J. (1762). *The Social Contract*.
65. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.
66. Scott, B. (2014). Visions of a techno-leviathan: The politics of the bitcoin blockchain. *E-International Relations*.
67. Sheridan, T. B. (1992). *Telerobotics, automation, and human supervisory control*. MIT press.
68. Sheridan, T. B. (2006). Supervisory control. *Handbook of Human Factors and Ergonomics, Third Edition*, pages 1025–1052.
69. Sigmund, K., De Silva, H., Traulsen, A., and Hauert, C. (2010). Social learning promotes institutions for governing the commons. *Nature*, 466(7308):861–863.
70. Skyrms, B. (2014). *Evolution of the social contract*. Cambridge University Press.
71. Standing Committee of the One Hundred Year Study of Artificial Intelligence (2016). Artificial intelligence and life in 2030. Technical report, Stanford University.
72. Sweeney, L. (2013). Discrimination in online ad delivery. *Queue*, 11(3):10.

73. Tambe, M., Scerri, P., and Pynadath, D. V. (2002). Adjustable autonomy for the real world. *Journal of Artificial Intelligence Research*, 17(1):171–228.
74. Thomaz, A. L. and Breazeal, C. (2008). Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*, 172(6):716–737.
75. Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly review of biology*, pages 35–57.
76. Tufekci, Z. (2015). Algorithmic harms beyond facebook and google: Emergent challenges of computational agency. *J. on Telecomm. & High Tech. L.*, 13:203.
77. Turchin, P. (2015). *Ultrasociety: How 10,000 Years of War Made Humans the Greatest Cooperators on Earth*. Beresta Books.
78. Valentino-DeVries, J., Singer-Vine, J., and Soltani, A. (2012). Websites vary prices, deals based on users’ information. *Wall Street Journal*, December 24.
79. Van de Poel, I. (2013). Translating values into design requirements. In *Philosophy and engineering: Reflections on practice, principles and process*, pages 253–266. Springer.
80. Young, H. P. (2001). *Individual strategy and social structure: An evolutionary theory of institutions*. Princeton University Press.