AI and the Social Contract

Chee Hae Chung[1], Daniel S. Schiff[1]
[1]Purdue University
chung382@purdue.edu, dschiff@purdue.edu

Abstract

As artificial intelligence (AI) systems increasingly shape public governance, they challenge foundational principles of political legitimacy. This paper evaluates AI governance against five canonical social contract theories—Hobbes, Locke, Rousseau, Rawls, and Nozick—while examining how structural features of AI strain these theories' durability. Using a structured comparative framework, the study applies three forms of legitimacy (procedural, moral-substantive, and recognitional) and three types of consent (explicit, tacit, and hypothetical) as normative benchmarks. Applying each theory, the analysis finds AI governance is marked by deficits in accountability, participation, rights protection, fairness, and freedom from coercion, while AI's opacity, global influence, and hybrid public-private control reveal blind spots within the social contract tradition itself. Though no single theory offers a complete solution and each contains specific weaknesses, the paper develops a hybrid model integrating Hobbesian accountability, Lockean rights protections, Rousseauian participation norms, Rawlsian fairness, and Nozickian safeguards against coercion. The paper concludes by distilling normative priorities for aligning governance with these hybrid contractarian standards: embedding participatory mechanisms, encouraging pluralistic ethical perspectives, ensuring institutional transparency, and strengthening democratic oversight. These interventions aim to reconfigure the social contract—and AI—for an era in which algorithmic systems increasingly mediate the exercise of political authority.

Introduction

"I still expect that it will be some change required to the social contract given how powerful we expect this technology to be...the whole structure of society itself will be up for some degree of debate and reconfiguration"
– Sam Altman

Artificial intelligence (AI) is increasingly embedded across public systems of governance, informing decisions related to welfare entitlements, policing, employment, and administrative coordination. Often developed by private companies and deployed with limited transparency or oversight, these systems raise fundamental questions about how authority is exercised and legitimated within democratic polities. Algorithmic systems now serve as structural mediators between the citizen and the state, shaping entitlements, civil freedoms, and opportunities for public participation (Crawford 2021; Eubanks 2018; Pasquale 2015). Certain structural features of AI and its governance make these legitimacy questions especially acute: their tendency to produce collective, diffuse harms that do not map neatly onto individual or product liability frameworks; responsibility gaps that emerge from opaque design and decision-making processes; and the mismatch between AI's well-established global reach and the territorially bounded authority of most governing bodies. These normative and structural challenges urge an assessment of the legitimacy of modern AI governance, particularly when those impacted possess little ability to comprehend, challenge, or assent to their deployment (Ananny and Crawford 2018; Mittelstadt 2019; Wagner 2016).

Social contract theory provides a foundational framework for evaluating the legitimacy of political authority. The tradition, spanning from classical philosophers like Hobbes, Locke, and Rousseau to contemporary theorists such as Rawls and Nozick, posits that governance is legitimate when institutions are based on consent, provide public goods, and safeguard liberty and fairness (Freeman 2007; Gaus 2010). Despite their divergent perspectives on sovereignty, justice, and rights, they largely converge on the idea that authority requires more than coercion or legal enforcement; rather, it must be justified to those who are subject to it. In the context of algorithmic governance, where individuals frequently lack participation opportunities, confront automated decisions impacting fundamental rights, and encounter systems devoid of meaningful contestation mechanisms, this tradition offers a powerful conceptual lens to assess the ethical implications of such emerging institutional arrangements (Beetham 2013; Tyler 2006).

While scholarship has explored AI through contractarian ideas, including concepts such as the "algorithmic social contract" (Rahwan 2018), "machine justice" (Gabriel 2022), and AI legitimacy within transnational regimes (Erman and Furendal 2024), most draw on a single philosophical tradition, often Rawlsian or constructivist. In contrast, the internal diversity of the social contract tradition—encompassing absolutist, republican, liberal, and libertarian strands—has not been applied in a systematic manner to assess AI's implications for the social contract. As a result, we lack a clear and comprehensive picture of how contemporary AI and its governance measure up to the political legitimacy standards articulated by different theories. Moreover, there is little understanding of how effectively these traditions can account for AI's structural features, or whether their underlying assumptions and proposals remain adequate in the algorithmic era.

This paper addresses these gaps by conducting a comparative analysis, applying five key social contract theories—Hobbes, Locke, Rousseau, Rawls, and Nozick. These canonical theorists were selected because their approaches represent distinct philosophical models: sovereign absolutism, natural-rights liberalism, civic republicanism, egalitarian constructivism, and libertarian minimalism, respectively (Freeman 2007; Gaus 2010). Each provides a different account of legitimate authority, and each has had substantial influence on liberal-democratic theory and contemporary governance. The analysis focuses on how each theorist defines two core mechanisms of political authority: subjective legitimacy and consent, dimensions that constitute key normative foundations of democratic governance (Habermas 2015), and that are under strain in the contemporary AI-mediated world.

Subjective legitimacy refers to a citizen's belief that a system of authority is justified, and is composed of perceptions surrounding fairness, recognition, and accountability. It reflects the alignment between institutional design and public moral standards (Beetham 2013; Tyler 2006). In the context of AI governance, such legitimacy is arguably weakened by opacity, limited public participation, and the difficulty of attributing responsibility for decisions, among other factors.

Consent, especially in its hypothetical forms, often operates as a moral test: would reasonable individuals, under impartial conditions, agree to the institutional rules that govern them (Habermas 2015; Rawls 1971)? Under current AI regimes, individuals typically do not grant explicit consent to data-driven policy decisions and rarely encounter mechanisms that meet the standards of democratic justification (Eubanks 2018; Srinivasan and Ghosh 2023), such as sufficient knowledge, meaningful access, and genuine prospects for contestation or remediation (Wagner 2016; Ananny and Crawford 2018; Mittelstadt 2019).

The comparative analysis in this study utilizes these two analytic axes, reviewing three forms of legitimacy—procedural, moral-substantive and recognitional, and three notions of consent—explicit, tacit, and hypothetical. These categories, grounded in established political theory scholarship (Gaus 2010; Riley 2013), serve as useful benchmarks for evaluating contemporary AI and its governance. The comparative method employed here allows for a systematic and transparent comparison of how each thinker justifies authority, defines the grounds of agreement, and sets the conditions for legitimate governance. The analysis reveals that algorithmic systems challenge these standards in distinct ways across theoretical models, exposing the points at which canonical requirements for legitimacy and consent are least compatible with current AI practices.

By mapping these requirements against the operational realities of AI governance, the study ultimately identifies a legitimacy gap that spans all traditions. While each theory sets its own justificatory thresholds, all five reveal incompatibilities with the current approach to AI in public governance. In some traditions, this misalignment concerns the absence of voluntary agreement; in others, the lack of moral justification or collective authorship. As such, this gap cannot be addressed through modest refinements to procedural compliance alone; instead, it reflects a deeper structural misalignment between technological authority and the normative conditions for political legitimacy that may urge a reformation of the social contract. In response, the study draws on insights from all five theories and proposes normative strategies to better calibrate legitimacy under AI governance. These include embedding participatory mechanisms, ensuring institutional transparency, and strengthening democratic oversight in ways that better reflect pluralistic perspectives (Helbing et al. 2019).

This study contributes to theoretical and practical debates. Theoretically, it advances an analytic framework that differentiates types of legitimacy and consent across five major contractarian scholars, revealing points of divergence and convergence raised by modern conditions and underexamined in the literature. Practically, it develops evaluative standards and normative insights that can guide the institutional design and democratic oversight of algorithmic systems. Two further notes are in order. First, while the analysis centers on certain canonical theories, critical traditions, including feminist and postcolonial critiques of the social contract, may further illuminate or challenge the limits of the canonical frameworks used here (Mills 2019; Okin 1989). Second, AI systems are not treated as moral agents themselves in our analysis, but rather understood as structural mediators of political power. Future work may benefit from conceptualizing AI systems as actors in the social contract.

The remainder of the paper proceeds as follows. It begins with a brief justification of the relevance of social contract theory for AI governance, and an introduction to the study's analytical focus on legitimacy and consent. This is followed

by a detailed review of how the canonical philosophers conceptualized the three notions of legitimacy and three notions of consent. Building on this comparison, the analysis then applies these categories to assess how structural features of AI interact with and ultimately strain various contractarian standards. The discussion then advances a hybrid normative model that incorporates the strengths of multiple traditions while addressing their weaknesses, offering concrete yet adaptable strategies for institutional design, transparency, and accountability in algorithmic systems. The paper concludes by reflecting on its theoretical and practical contributions, noting its limitations, and outlining future research directions at the intersection of political theory and the governance of emerging technologies, towards the preservation (or reconfiguration) of the social contract.

Why Social Contract Theory for AI Governance?

Numerous philosophical and ethical lenses have been applied to make sense of AI's implications, including applied, systematic/normative, and metaethical approaches. Yet this study's interest in social contract theory is motivated by its normative focus on the legitimacy of institutional authority. Contractarian theory is well-positioned to make sense of these issues which are underexamined compared to more practical, day-to-day issues raised regularly in applied ethics debates (e.g., surrounding privacy or algorithmic bias in a certain context). For instance, while consequentialist ethics evaluates authority by outcome utility and virtue ethics emphasizes individual moral cultivation (Nussbaum 2011), social contract theory offers a distinct advantage in addressing AI's governance dilemmas: it evaluates whether the rules structuring authority are morally justified.

This emphasis is especially relevant for automated decision-making systems, where rule-based authority is often exercised without transparent deliberation or interpersonal accountability. As Pettit (2012) urges, legitimacy requires more than procedural regularity; it demands that institutional actions can be justified to those subject to them. In algorithmic environments, however, decisions often arise from opaque optimization or private datasets, placing them beyond the justificatory reach of affected citizens. Most individuals lack both epistemic access and procedural standing in AI-driven infrastructures, rendering conventional accountability mechanisms ineffective (Srinivasan and Ghosh 2023). For example, algorithmic immigration screening tools have denied entry to applicants based on opaque risk models, without offering those affected a meaningful opportunity to understand or contest the decision (Molnar and Gill 2018).

Given such challenges, social contract theory offers a framework not only for asking whether power is effectively exercised, but whether it should be exercised in the first place. Its principles of mutual justification, rational agreement, and reciprocal duty make it especially relevant for examining AI systems that increasingly make policy-relevant determinations without clear channels for democratic validation or redress and for public governance of AI broadly, including within the private sector.

Core Dimensions: Legitimacy and Consent

Within the social contract tradition, two justificatory mechanisms are especially central to establishing (and assessing) the moral authority of political institutions: subjective legitimacy and consent. These mechanisms form the normative foundation for claims to political obligation in liberal-democratic theory (Habermas 2015).

Subjective legitimacy concerns the normative justification of authority—whether it is rightfully exercised and recognized by those subject to it (Beetham 2013; Buchanan 2002). Political theory commonly distinguishes three forms: procedural, moral-substantive, and recognitional legitimacy (Gaus 2010; Riley 2013). In other words, legitimacy holds a tripartite structure: legality (conformity with rules), shared belief in legitimacy (normative endorsement), and consent (expressed or implied). Empirical studies show that compliance is sustained when citizens perceive institutions as procedurally fair and morally grounded (Tyler 2006).

Consent operates as a moral filter on the use of power: are the governed in a position to reasonably accept the rules that bind them? Classical forms of consent include explicit contracting (Hobbes), tacit obedience (Locke), and active participation (Rousseau), while contemporary theory favors hypothetical consent (Rawls).

In AI-governed contexts, both mechanisms face challenges. Decisions and institutional designs are often opaque, unaccountable, and imposed without meaningful deliberation. Consent in digital contexts is equally lacking, often defaulting into passive data extraction advanced by platform logic, lacking the reflective endorsement envisioned by democratic theory. These developments raise pressing questions: Can AI systems meet the standards of political authority? If not, which dimensions of traditional legitimacy are violated, and how might they be restored?

Divergent Traditions of Legitimacy and Consent in Classical Social Contract Theories

Legitimacy Dimensions

Procedural Legitimacy refers to the extent to which authority is grounded in adherence to formally established rules, institutional roles, and decision-making procedures. In Hobbes's framework, procedural legitimacy is tied to the

sovereign's ability to enforce laws and maintain order. In *Leviathan* (1651), Hobbes declares, "covenants, without the sword, are but words and of no strength" (ch. XVII), indicating that legitimacy stems from the sovereign's effective capacity to command rather than from public recognition or moral agreement. As Skinner (2008) observes, Hobbes seeks to relocate all notions of obligation within the framework of law as laid down by sovereign power. This model leaves little space for procedural fairness beyond compliance with enacted rules, strongly prioritizing stability over participatory mechanisms.

In Locke's account, procedural legitimacy also depends on rule conformity but is conditional upon the protection of natural rights. Legislative authority must be exercised "by promulgated standing laws" and "not by extemporary decrees" (*Second Treatise*, §124). Laws must be general, known, and prospective to be legitimate, embedding procedural requirements within a broader moral framework. Tuckness (2002) notes that such conditionality constrains arbitrary state action and offers safeguards absent in Hobbesian absolutism, linking procedural formality to the moral purpose of protecting liberty and property.

Rousseau integrates procedural legitimacy through the enactment of laws that reflect the general will, requiring that "the people, being subject to the laws, ought to be their author" (*Social Contract*, Book II). This embeds procedure within participatory self-legislation. Yet it remains vulnerable to the critique that large or technologically mediated societies may lack mechanisms to ensure that formal legislative processes truly track the general will (Berlin 2014; Eubanks 2018; Srinivasan and Ghosh 2023).

Rawls reframes procedural legitimacy through the concept of "pure procedural justice," where fairness in decision-making ensures the justice of outcomes only if background conditions are just (Rawls 1971). This vision presupposes equality of opportunity and impartiality, conditions often absent in algorithmic governance, where opaque systems can perpetuate structural disadvantage (Mills 2019). Algorithmic credit scoring, for example, can disproportionately lower ratings for historically disadvantaged groups, thereby undermining fair equality of opportunity even if the procedure is technically consistent.

Nozick gives procedural legitimacy minimal scope beyond its role in enforcing voluntary agreements. In *Anarchy, State, and Utopia* (1974), legitimacy is satisfied if transactions respect historical entitlements, regardless of distributive outcomes. This narrow proceduralism is ill-suited for AI governance contexts in which harms emerge from structural features, such as pervasive data aggregation, occurring without meaningful individual transactions much less explicit agreement. Data brokerage ecosystems, which aggregate personal information without explicit user engagement, would likely fall outside Nozick's procedural concern, even when they lead to large-scale profiling.

Moral-Substantive Legitimacy refers to the extent to which authority is justified by its adherence to ethical standards such as fairness, justice, rights protection, or the promotion of the common good, rather than by procedural compliance alone (Beetham 2013). In Hobbes's framework, moral-substantive considerations are secondary to stability and order. For Hobbes, the sovereign's role is not to realize distributive justice but to maintain peace, since "where no common power, there is no law; where no law, no injustice" (*Leviathan*, ch. XIII). This places Hobbesian legitimacy largely outside moral-substantive tests in the modern democratic sense, as the sovereign's actions are evaluated in minimalistic terms of preventing societal collapse, not aligning with shared ethical norms (Kavka 2021).

Locke, by contrast, embeds moral-substantive legitimacy explicitly in the preservation of natural rights, including life, liberty, and property, as the central end of political society. Authority becomes illegitimate when it fails to protect these rights, even if procedural norms are followed. In Locke's *Second Treatise* (§222), he argues that when legislators violate natural rights, they "put themselves into a state of war with the people," thus forfeiting legitimacy (Simmons 2000). However, Locke's framework assumes a clear and stable definition of natural rights, which can be contested or selectively interpreted in pluralistic, technologically mediated societies, limiting its applicability and explanatory power in complex governance environments.

Rousseau's moral-substantive legitimacy derives from alignment with the general will, understood as the collective moral orientation toward the common good. Laws are legitimate when they embody this will, which Rousseau distinguishes from the mere aggregation of private interests (*Social Contract*, Book II). In this view, legitimacy is lost when laws privilege sectional interests over the common good. However, the ambiguity of the general will creates risks of coercion, where dissent may be suppressed under the pretext of pursuing collective morality (Berlin 2014).

Rawls defines moral-substantive legitimacy through the two principles of justice articulated in the original position (described later): the equal basic liberties principle and the difference principle. Legitimacy depends on whether institutions respect these principles and maintain fair equality of opportunity (Rawls 1971). Rawls explicitly links legitimacy to distributive fairness, arguing that departures from just arrangements require unanimous consent of the disadvantaged parties. However, Rawls's framework, grounded in ideal theory, has difficulty engaging with entrenched injustices and the systemic biases of real institutions, which are amplified and complicated in opaque AI-mediated contexts.

Nozick rejects patterned distributive principles as a basis for legitimacy, grounding moral-substantive authority in historical entitlement and the justice of acquisition and

transfer. If holdings arise from just processes, the resulting distribution is legitimate regardless of inequality (Nozick 1974). While this view offers a clear standard for evaluating coercion, it sidelines questions of systemic fairness and collective welfare, rendering it poorly equipped to assess legitimacy in AI governance contexts where benefits and burdens are distributed through opaque, non-consensual systems that shape structural opportunities (O'Neil 2017).

Recognitional Legitimacy concerns whether authority is affirmed and accepted by those subject to it, not only in terms of compliance, but also with respect to moral and symbolic acknowledgement of its right to govern. It reflects the degree to which citizens perceive institutions as representing their values, identity, and political community.

In Hobbes's model, recognitional legitimacy plays a minimal role. Because authority derives from the sovereign's capacity to enforce peace, Hobbes sees no requirement for citizens to identify with or morally endorse the sovereign's decisions. As he states in Leviathan, "the obligation of subjects to the sovereign lasts as long, and no longer, than the power lasteth" (ch. XXI). Skinner (2008) emphasizes that Hobbes deliberately sidelines mutual recognition in favor of a security-based compact, leaving little scope for or interest in evaluating governance through shared symbolic commitment.

Locke's theory, by contrast, treats recognitional legitimacy as integral to sustaining political society. Because authority rests on consent and trust, citizens must continue to view the government as a rightful trustee of their rights. The withdrawal of this recognition, when rulers "endeavor to take away and destroy the property of the people" (*Second Treatise*, §222), justifies the dissolution of government. Locke's emphasis on trust implies that recognitional legitimacy is fragile in contexts of systemic opacity or unaccountable decision-making (Tuckness 2009), both of which are endemic to AI governance.

For Rousseau, recognitional legitimacy is foundational: citizens must see themselves as both authors and subjects of the laws. This civic identity is what turns obedience into self-rule. In *Social Contract* (Book II), Rousseau insists that laws are legitimate only when they arise from collective authorship expressing the general will. However, this model assumes a high degree of civic unity and participation, which may be infeasible in large-scale, technologically mediated polities where algorithmic decision-making is removed from public deliberation (Pateman 2015).

Rawls incorporates recognitional legitimacy through the idea of public reason: institutions are legitimate when their exercise of political power can be justified to all citizens as free and equal, using reasons they can reasonably accept (Rawls 1993). This requirement entails that legitimacy is sustained not merely through just outcomes but also through ongoing mutual recognition of shared political membership. Yet as Habermas (2015) points out, this standard presupposes transparent deliberation and accessible justificatory processes, both of which are often disrupted by AI's opaque logic and common conditions of private control.

Nozick assigns minimal importance to recognitional legitimacy, as his framework does not require citizens to identify with the state or its principles beyond refraining from coercion. Legitimacy, in his account, does not depend on mutual recognition but on the state's non-violation of individual rights (Nozick 1974). This leaves little normative space for addressing legitimacy crises that stem from public alienation or distrust, such as issues that are central to debates over AI governance in public administration (Srinivasan and Ghosh 2023).

Consent Dimensions

Explicit Consent refers to a clear, deliberate, and documented agreement by individuals to submit to political authority or specific rules. In social contract theory, it represents the most unambiguous form of legitimating assent, though its feasibility and prevalence have been widely debated (Simmons 2014). Hobbes frames the commonwealth's founding as a collective covenant where each person authorizes the sovereign: "I authorize and give up my right of governing myself, to this man, or to this assembly of men" (*Leviathan,* ch. XVII). This act is explicit in the initial contract, but thereafter, continued legitimacy does not depend on ongoing consent—only on the sovereign's capacity to provide security. Hobbes treats explicit consent as a one-time founding event (Hampton 1986), which poses a challenge for algorithmic governance that evolves over time without renewed public endorsement.

Locke assigns greater importance to explicit consent, particularly in situations where individuals voluntarily join a political society. In *Second Treatise* (§119), he writes, "no one can be subjected to the political power of another, without his own consent." This form of consent is rare in practice, as Locke acknowledges, but it functions as the clearest form of legitimate authority. For AI governance, explicit consent might parallel formal opt-in mechanisms for algorithmic decision-making. However, real-world AI deployments rarely allow citizens to grant or withhold consent in this clear form, undermining the Lockean ideal.

Rousseau's conception of explicit consent is bound up with active participation in lawmaking. For him, each citizen's vote on legislation constitutes an explicit act of consent to the general will (*Social Contract*, Book II). While this provides a robust participatory foundation, scholars have pointed out that modern governance rarely affords citizens direct legislative input (Berlin 2014; Habermas 2015), including in AI-driven administrative contexts where policy is even more distantly embedded in technical code.

Tacit Consent refers to the implied agreement to political authority inferred from individuals' actions or circumstances, such as residence within a polity, acceptance of its protection, or participation in its institutions (Simmons 1993). Although more flexible than explicit consent, it raises enduring questions about voluntariness, awareness, and the possibility of meaningful dissent.

Hobbes suggests continued obedience to the sovereign constitutes acceptance of their authority, even if initial consent was explicit (*Leviathan*, ch. XXI). However, in his framework, survival imperatives effectively compel compliance, making tacit consent indistinguishable from submission under necessity. This conflation strips tacit consent of normative force (Pettit 2012), as in automated welfare adjudication where refusal means material deprivation. This limitation highlights Hobbes's tendency to conflate consent with mere endurance under authority.

Locke offers the most systematic account of tacit consent in *Second Treatise of Government* (§119–122). He argues that individuals give tacit consent "by actually enjoying any of the dominions of any government," such as using its infrastructure or holding property under its laws. This form of consent is revocable, and individuals retain the right to exit the polity if dissatisfied (§121). However, Locke's formulation presupposes realistic alternatives to remaining in the polity, a condition rarely met in practice, as citizens often have no viable opt-out option, thereby undermining the voluntariness central to Lockean tacit consent (Simmons 2000; Tuckness 2009).

Rousseau is more skeptical of tacit consent as a legitimating mechanism. In *The Social Contract*, he argues that legitimate political authority must rest on active, ongoing participation in expressing the general will (Book IV). Passive acquiescence or failure to dissent does not constitute consent, as it lacks deliberate political engagement. This view limits Rousseau's framework in large, complex societies where continuous, meaningful participation is structurally difficult.

Rawls does not frame tacit consent as a central pillar of legitimacy, but he does acknowledge that established practices often persist through implicit public acceptance. His constructivist approach evaluates such practices by whether they could be justified under fair conditions, rather than by their factual acceptance (Rawls 1971). It shows that his framework grants less normative weight to tacit consent that is sustained by habit rather than reasoned agreement.

Nozick's minimal state theory implicitly relies on a form of tacit consent through non-resistance to protective associations. Yet, without real opportunities for dissent or withdrawal, such non-resistance may be indistinguishable from imposed compliance (Cohen 2012). This reveals the

Theorist	Procedural Legitimacy	Moral-Substantive Legitimacy	Recognitional Legitimacy	Explicit Consent	Tacit Consent	Hypothetical Consent
Hobbes	Legitimacy derives from sovereign compliance with established laws ensuring peace and security.	Stability prioritized over distributive justice; sovereign not evaluated by moral-substantive tests beyond preventing disorder.	Minimal role; legitimacy rests on power maintenance rather than shared identification.	Founding covenant in which subjects permanently authorize the sovereign.	Continued obedience signals acceptance, often from necessity.	Proto-hypothetical consent based on rational fear of insecurity in the state of nature.
Locke	Laws must be general, known, and prospective, protecting natural rights.	Authority exists to preserve life, liberty, and property.	Trust-based recognition; rulers as rights trustees.	Voluntary joining of political society.	Use of public goods implies revocable consent, with the right of exit.	Limited: reasonable individuals would accept rights-protecting authority.
Rousseau	Laws reflect the general will through citizen authorship.	The common good as the moral standard for authority.	Foundational; self-rule as co-authorship of laws.	Legislative participation constitutes explicit consent.	Rejected as insufficient for legitimacy.	Equal participation assumed in the general will.
Rawls	Fair procedures legitimate outcomes under just background conditions.	Justice as fairness: equal basic liberties and the difference principle.	Public reason ensures mutual recognition among free and equal citizens.	Counterfactual consent via the original position and veil of ignorance.	Marginal role in legitimacy.	Central device for legitimacy: impartial choice in the original position.
Nozick	Enforcing voluntary agreements and protecting rights.	Historical entitlement, not distributive patterns, as the legitimacy basis.	Minimal role; recognition beyond rights compliance unnecessary.	Explicit voluntary association.	Non-resistance to protective associations as tacit consent.	Rejected as a basis for legitimacy.

Table 1: Legitimacy and Consent in Canonical Social Contract Theories

limits of his account in settings where individuals cannot meaningfully avoid or reject governance arrangements.

Hypothetical Consent evaluates the legitimacy of political authority by asking whether reasonable individuals, under specified impartial conditions, would agree to the governing arrangements in question. This device is counterfactual and normative rather than empirical, seeking to derive legitimacy from the reasonableness of imagined agreement rather than actual assent.

Hobbes could be read as invoking a proto-form of hypothetical consent in his argument that rational individuals, in a pre-political "state of nature," would agree to submit to an absolute sovereign to escape the "war of all against all" (*Leviathan*, ch. XIII). Yet Hobbes's conditions are so shaped by fear and necessity that they collapse the normative distinction between reasonable agreement and compelled submission (Skinner 2008).

Locke is more ambivalent about hypothetical consent. While he appeals to natural law as binding on all rational creatures, he grounds legitimacy primarily in actual or tacit consent. Nonetheless, Locke occasionally implies that reasonable individuals would consent to governments that protect natural rights, bringing him into limited alignment with hypothetical consent logic (Simmons 2014). This conditional acceptance means Locke's hypothetical consent applies only to rights-preserving arrangements, limiting its scope in broader governance contexts.

Rousseau's conception of the "general will" shares a kinship with hypothetical consent insofar as it embodies the collective will of citizens directed toward the common good, which they would endorse under conditions of equal participation. However, Rousseau's model is less about counterfactual agreement and more about actualized civic authorship. Still, both frameworks rely on the assumption that citizens can deliberate from a standpoint of equality, a condition that could be easily challenged when real-life processes exclude most citizens from meaningful engagement or understanding (Pateman 2015).

Rawls offers the most influential formulation through the "original position" and "veil of ignorance" in *A Theory of Justice*. In this thought experiment, individuals stripped of knowledge about their personal characteristics, social status, and natural talents would select principles of justice that maximize fairness and protect the least advantaged (Rawls 1971). Hypothetical consent, here, legitimizes institutions that would be chosen in such fair conditions, regardless of actual consent in the real world. However, critics such as Mills (2019) contend that Rawls's idealization risks ignoring entrenched social and technological inequalities.

Nozick explicitly rejects hypothetical consent as a basis for legitimacy. In *Anarchy, State, and Utopia* (1974), he argues that the fact that people would agree to certain arrangements under imagined conditions does not grant moral permission to impose those arrangements in reality. Legitimacy, for him, rests on respecting actual entitlements and voluntary associations. This categorical dismissal excludes hypothetical consent as a basis for legitimacy in his theory.

Theoretical Divergences and Limitations in Social Contract Theories

Despite their foundational influence, these canonical social contract theories exhibit conceptual divergences and internal limitations that complicate their ability to guide contemporary algorithmic governance. These limits appear as both theoretical gaps and normative blind spots in addressing authority that is neither wholly public nor procedurally participatory. Comparative analysis of the five thinkers reveals tensions in their treatments of legitimacy and consent, shaping how their frameworks can or cannot accommodate AI-based decision-making institutions.

First, there is an unresolved disagreement among theorists regarding the very source of legitimacy. For Hobbes, legitimacy derives from institutional efficacy in preventing conflict and preserving order, where "the value or worth of a man is, as of all other things, his price" (*Leviathan*, ch. X). His model suppresses any need for recognition-based or moral-substantive legitimacy, which renders it ill-equipped for evaluating AI systems whose harms are neither violent nor visible but operate through opacity and statistical profiling (Eubanks 2018). For example, a Hobbesian government could justify the continuous use of AI-powered risk assessment tools in policing purely on their capacity to lower crime rates, even if they disproportionately target certain communities and erode trust. As Flathman (1993) argues, Hobbes offers a model of authority that is unresponsive to the plural and interpretive moral claims essential to modern democratic legitimacy.

In contrast, Locke emphasizes the importance of tacit and revocable consent based on the preservation of natural rights. However, this reliance on tacit agreement becomes increasingly problematic in digital contexts. Locke's logic assumes a feasible "right of exit," which collapses under conditions of platform dependence and infrastructural surveillance. When citizens cannot functionally participate in society without using AI-mediated public services, such as automated tax filing, digital identity verification, or algorithmic welfare distribution, the possibility of meaningful exit is effectively nullified. The mere act of residing in a jurisdiction or using a digital service is insufficient as a moral basis for binding authority, especially when users have little choice but to comply with embedded algorithmic decisions. Moreover, Locke's focus on individual property rights as the foundation of political obligation does not anticipate data-based economies, where ownership and agency are frequently decoupled (Cohen 2012). For instance, citizens may have legal ownership of personal data in principle but lack any practical control over how AI systems aggregate and process it.

Rousseau's participatory republicanism introduces a more demanding model of legitimacy grounded in collective authorship. His claim that "obedience to the law one prescribes to oneself is freedom" (*The Social Contract*, Book I) sets a high normative standard. Yet, this conception is vulnerable to coercive reinterpretation, particularly when "the general will" is operationalized without direct participation or transparency (Berlin 2014). Algorithmic policymaking in environmental regulation, for example, might be defended as serving the collective good but still bypass public input entirely by embedding decisions in technical systems inaccessible to ordinary citizens. In algorithmic regimes, where citizens are neither aware of decision-making logics nor consulted in their formulation, the very possibility of being a co-author of the rules becomes illusory. Contemporary platforms reverse Rousseau's aspiration by automating rule-setting in ways that bypass civic engagement entirely (Zuboff 2019).

Rawls's theory of justice, while offering a sophisticated model of hypothetical consent, has also drawn criticism for its abstraction. The veil of ignorance presumes epistemic symmetry and shared moral reasoning, assumptions strained by AI's hidden architectures and asymmetric information flows (Mills 2019). For example, predictive analytics in healthcare may allocate resources based on biased training data, producing outcomes that those behind the veil of ignorance could not rationally endorse if they understood the biases; yet in practice, most stakeholders lack both the access and expertise to evaluate these systems and their complex downstream impacts. Further, this approach does not fully account for how algorithmic personalization fragments the public sphere, undermining the plausibility of shared reasoning upon which his constructivist legitimacy depends (Citron and Pasquale 2014).

Finally, Nozick's libertarian minimalism dismisses the need for consent altogether once property rights are respected. His claim that redistributive taxation is equivalent to "forced labor" offers a radically individualist foundation for legitimacy that struggles to accommodate the collective and relational nature of digital infrastructures. In practice, platforms centralize control over user data while diffusing accountability, a form of "infrastructural power" that eludes Nozick's narrow conception of coercion (Cohen 2012). For instance, even if a privately-run AI infrastructure does not explicitly violate property rights, it could still wield enormous *de facto* power over access to jobs, credit, and public participation, effects that Nozick's framework is arguably poorly equipped to evaluate.

What unites these frameworks, despite their divergences, is their limited anticipation of authority that emerges from hybrid, non-state actors, operates across borders, and governs through probabilistic and opaque logics. Modern AI systems, particularly those owned and operated by multinational technology firms, often perform quasi-sovereign functions such as identity verification, credit scoring, and information curation, all without being subject to the institutional legitimacy tests designed for traditional state actors. As Rahwan (2018) notes, algorithmic systems introduce a "machine behavior" paradigm that cuts across traditional political boundaries and ethical theories.

Moreover, the reach of the private sector altogether, including its dominance in AI design, implementation, and evaluation, likewise undermines the robustness of any overly state-centric notion of governance. Despite their significant contributions then, none of the five thinkers meaningfully engages with (or anticipates) the implications of sustaining political legitimacy within a governance model where authority is often exercised without visibility, representation, direct accountability, or even state control.

This comparative analysis highlights that while the selected theorists offer valuable tools for conceptualizing legitimacy and consent, their models require reinterpretation or supplementation to remain normatively sufficient. Their limitations suggest that AI governance cannot be assessed using these inherited categories alone. Instead, it calls for a reconfiguration of contractarian principles to reflect the structural shifts introduced by computational institutions, especially in relation to consent that is rarely explicitly sought and legitimacy that is not publicly deliberated.

Discussion: Implications and Normative Strategies for AI Governance

The comparative analysis reveals that while classical social contract theories offer diverse and valuable frameworks for assessing legitimacy, none of them, in their original form, can fully accommodate the structural and epistemic challenges posed by AI governance. This finding has two major implications: (1) that legitimacy deficits in algorithmic governance are not incidental but structurally embedded in the technology's design, deployment, and control; and (2) that recalibrating legitimacy will require hybridizing and extending the principles of social contract theory to address these deficits.

Theorizing Legitimacy for the AI Age

The first implication concerns the *nature* of legitimacy in algorithmic contexts. In social contract theory, legitimacy rests on a combination of consent, moral justification, and recognition. However, AI systems often bypass each of these mechanisms: they rarely solicit explicit or tacit consent in meaningful ways; their inner workings are inaccessible, limiting moral evaluation; and they operate invisibly, eroding public recognition of their authority. This suggests that traditional legitimacy tests may need to be reframed in at least three ways:

First, subjective legitimacy should be prioritized, even operationalized as an empirical and dynamic measure, i.e., tracking whether affected communities *believe* algorithmic governance is fair, transparent, and accountable. For

AI Governance Dilemma	Limitation of Canonical Theories	Hybrid Principles
Collective Harms vs. Individual Liability	**Hobbes**: Oriented toward maintaining order, with limited mechanisms for addressing dispersed or cumulative harms. **Locke**: Tacit consent strained in platform-dependence with limited exit options. **Rousseau**: Participatory authorship challenged in large-scale, tech-mediated rule-making **Rawls**: Relies on assumptions of transparency and informational symmetry that may not hold in complex AI systems. **Nozick**: Lacks tools for addressing collective harms absent direct rights violations.	**Hybrid Principle of Distributed Accountability:** Clear attribution of responsibility (Hobbes) + reversible rights-affecting decisions (Locke) + fairness audits (Rawls) + safeguards against coercion (Nozick).
Responsibility Gap from Opacity	**Hobbes**: Does not incorporate requirements for institutional transparency. **Locke**: Public justification principles undermined when decision-making is not accessible to public understanding. **Rousseau**: Co-authorship norms inapplicable where interpretability is absent. **Rawls**: Public reason is limited when opacity obstructs reciprocal justification. **Nozick**: Accepts non-transparency where no explicit rights violations are evident.	**Transparency–Recognition Principle:** Rawlsian public reason + Rousseauian co-authorship, operationalized via explainability mandates and public-facing oversight dashboards.
Global Reach vs. National Governance	**Hobbes**: Sovereignty framework bounded by territorial jurisdiction. **Locke**: Rights protection primarily conceived within national legal frameworks. **Rousseau**: General will principles designed for nationally bounded polities. **Rawls**: Original position and fairness theory predominantly state-centric. **Nozick**: Provides limited resources for coordinated transnational governance.	**Polycentric Legitimacy Framework:** Integrate local participation (Rousseau), national accountability (Locke), and transnational ethics councils (Rawls) for AI oversight.

Table 2: Hybrid Principles to Address AI Governance Dilemmas

instance, national health systems using AI for patient triage could incorporate periodic citizen surveys, public audits, and community consultations to measure and adjust perceived legitimacy, rather than relying solely on expert assessments of system accuracy or risk acceptance.

Second, consent will need to be reconceptualized for AI-related environments where opting out is practically infeasible. This may require substituting the *opportunity for explicit consent* with *robust procedural safeguards* that approximate the conditions under which consent could be given. For example, AI-powered tax compliance systems could include deliberative opt-in windows for major policy changes, allowing citizens to review and challenge new algorithmic rules before implementation.

Third, recognitional legitimacy must be rebuilt through participatory and symbolic inclusion, ensuring that affected citizens identify with the governance process, not merely accept it as inevitable. Along these lines, public-facing dashboards for algorithmic policing or welfare systems could display real-time oversight metrics and community feedback channels, reinforcing the sense that these systems are "of" and "for" the community.

Hybridizing Social Contract Traditions

A second implication is that different strands of social contract theory offer complementary resources for AI governance. No single framework can address the multi-layered legitimacy gaps AI introduces. However, a hybrid approach could integrate their respective strengths:

- From Hobbes: the recognition that legitimacy requires a clear attribution of responsibility for decision outcomes, even when mediated by complex systems. This could support the principle that every AI system in public governance must have a legally accountable authority clearly named.
- From Locke: the insistence on revocability and the protection of fundamental rights, guiding the design of algorithmic systems with *built-in reversibility* of decisions where rights are at stake.
- From Rousseau: the requirement for participatory co-authorship, adapted for digital contexts via citizen assemblies, participatory algorithm design workshops, or crowd-sourced policy inputs.
- From Rawls: the fairness principle as a standard for algorithmic outputs, ensuring that AI systems are tested for disparate impact and, critically, adjusted to meet distributive justice criteria.
- From Nozick: the minimal coercion principle, reframed to require that citizens retain practical alternatives when interacting with AI systems, especially in contexts such as credit scoring, hiring, and identity verification.
- Finally, the recalibration of legitimacy should be informed by critical traditions often absent from canonical contract theory, including postcolonial, and intersectional critiques. These perspectives highlight, for instance, how many "universal" principles continue to mask structural exclusion and historical injustice.

Strategies for Restoring Legitimacy

Drawing on these hybridized insights, four concrete strategies emerge for recalibrating legitimacy under AI governance:

1. Participatory design and deliberation: AI governance processes should be opened to public participation at design, deployment, and evaluation stages. This might include *citizen juries* for reviewing high-stakes AI systems, participatory prototyping sessions, and open deliberation on ethical trade-offs. For instance, before rolling out an AI welfare eligibility tool, a government could host multi-stakeholder workshops including recipients, caseworkers, and technologists to jointly review algorithmic decision rules.
2. Institutional transparency and algorithmic audits: Legitimacy depends on the capacity for public and expert scrutiny. Mandatory independent audits, explainability standards, and public disclosure of system objectives and limitations are crucial. For example, predictive policing algorithms could be subject to quarterly independent bias audits, with results made available in public reports.
3. Democratic oversight and multi-level accountability: Given the transnational nature of many AI systems, oversight mechanisms should combine local, national, and international layers. This could involve transnational ethics councils and local watchdog committees working in conjunction with legislative oversight panels. For example, an international registry of AI systems deployed in public governance could require states to document their purposes, oversight mechanisms, and audit outcomes.
4. Pluralistic ethical integration: Since AI governance affects diverse communities, legitimacy should reflect multiple moral perspectives. This means integrating constructivist consent requirements (public justifiability) with fairness tests, rights protections, and minimal coercion safeguards. For instance, an AI-based immigration decision system could be evaluated against both Rawlsian fairness principles and Lockean rights protections, assessing acceptability across different moral doctrines.

Conclusion and Future Directions

This paper sets out to examine whether extant approaches to the social contract can stretch to account for the new challenges imposed by AI and the need to govern it. By developing and applying a structured comparative typology of legitimacy (procedural, moral-substantive, recognitional) and consent (explicit, tacit, hypothetical, constructivist), the analysis revealed that while each of the five examined thinkers, including Hobbes, Locke, Rousseau, Rawls, and Nozick, offers distinctive insights, none of these canonical approaches can, in isolation, address the multi-layered legitimacy deficits that AI systems introduce.

The findings demonstrate that these deficits are not accidental but structural. They emerge from the opacity, scale, and cross-jurisdictional nature of AI systems, their embedding in private-sector infrastructures, and the practical impossibility for citizens to meaningfully opt out. While Hobbesian, Lockean, Rousseauian, Rawlsian, and Nozickian frameworks each capture certain aspects of legitimacy, they leave critical blind spots: Hobbes underestimates moral and recognitional legitimacy, Locke overestimates the feasibility of revocable consent, Rousseau demands participatory structures rarely present in digital governance, Rawls assumes fairness can be evaluated under informational symmetry, and Nozick minimizes the importance of collective oversight.

The normative strategies proposed here are intended to recalibrate legitimacy for AI governance by hybridizing the strengths of these traditions while compensating for their respective weaknesses. This hybrid framework reframes subjective legitimacy as an ongoing, measurable public perception; reconceptualizes consent for contexts where opting out is unrealistic; and grounds recognitional legitimacy in participatory and symbolic inclusion. Three related avenues for future research are especially urgent:

1. Operationalizing legitimacy: Future work should focus on centering legitimacy normatively as well as developing empirical instruments to measure subjective legitimacy in AI governance contexts. This includes survey-based perception indices, participatory audit logs, and legitimacy dashboards that track trust, fairness, and recognition over time.
2. Embedding consent across AI: Research is needed on the technical and institutional mechanisms by which forms of explicit, hypothetical, and constructivist consent can be embedded into AI systems and their governance. For example, dynamic consent protocols could allow citizens to opt in or out of specific data uses over the lifecycle of an AI system.
3. Integrating critical perspectives: Canonical contractarian approaches should be supplemented with feminist, postcolonial, and intersectional analyses to capture how AI governance interacts with existing power asymmetries.

In conclusion, the social contract remains a valuable though incomplete lens for interrogating AI governance. Its value lies not in offering a single definitive legitimacy test but in providing a *plural set of principles* that, when adapted and combined, can illuminate the normative gaps in current governance structures. As AI continues to reshape the conditions under which political authority is exercised, the social contract must itself be renegotiated—expanded to encompass not only the relationship between citizens and the state, but also the algorithmic intermediaries that increasingly mediate that relationship.

Acknowledgements

We are grateful to Molly Scudder for her detailed and constructive feedback during the Purdue Political Science Research Workshop, which substantially improved the structure and clarity of this paper. We also thank Annette Zimmermann for her insightful comments during the Purdue AI Summit 2025, which helped refine the paper's theoretical framing, as well as Kaylyn Jackson Schiff and J.P. Messina for their valuable perspectives shared at the same event. We further appreciate Jo Ann Oravec for her thoughtful engagement as a discussant at the 2024 Midwest Political Science Association conference, which informed several revisions to the argument.

References

Abizadeh, A. 2016. Sovereign Jurisdiction, Territorial Rights, and Membership in Hobbes. In *The Oxford Handbook of Hobbes*, 397–431. Oxford: Oxford University Press.

Almeida, P. G. R.; C. D. dos Santos; and J. S. Farias. 2021. Artificial Intelligence Regulation: A Framework for Governance. *Ethics and Information Technology*, 23(3): 505–525.

Ananny, M.; and K. Crawford. 2018. Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability. *New Media & Society*, 20(3): 973–989.

Baldwin, R.; M. Cave; and M. Lodge. 2011. *Understanding Regulation: Theory, Strategy, and Practice*. Oxford University Press.

Banerjee, S.; P. K. Singh; and J. Bajpai. 2018. A Comparative Study on Decision-Making Capability Between Human and Artificial Intelligence. In *Nature Inspired Computing*, Singapore: Springer. 203–210.

Beetham, D. 2013. *The Legitimation of Power*. Bloomsbury Publishing.

Bergen, M. 2018. Google Grapples With 'Horrifying' Reaction to Uncanny AI Tech. *Bloomberg*.

Berlin, I. 1990. *Four Essays on Liberty*. Oxford University Press.

Berlin, I. 2014. Two Concepts of Liberty. In *Reading Political Philosophy*, Routledge. 231–237.

Bertram, C. 2012. Rousseau. In *The Routledge Companion to Social and Political Philosophy*, Routledge. 82–91.

Blue, G.; and D. Davidson. 2020. Advancing a Transformative Social Contract for the Environmental Sciences: From Public Engagement to Justice. *Environmental Research Letters*, 15(11): 115008.

Boucher, D.; and P. Kelly. 2003. *The Social Contract from Hobbes to Rawls*. Routledge.

Buchanan, A. 2002. Political Legitimacy and Democracy. *Ethics*, 112(4): 689–719.

Buchanan, J. 1975. *The Limits of Liberty: Between Anarchy and Leviathan*. University of Chicago Press.

Chinen, M. 2023. *The International Governance of Artificial Intelligence*. Edward Elgar Publishing.

Citron, D. K.; and F. Pasquale. 2014. The Scored Society: Due Process for Automated Predictions. *Washington Law Review*, 89: 1–33.

Coeckelbergh, M. 2020. *AI Ethics*. The MIT Press Essential Knowledge Series.

Cohen, J. E. 2012. *Configuring the Networked Self: Law, Code, and the Play of Everyday Practice*. Yale University Press.

Cranston, M. 1991. *Jean-Jacques: The Early Life and Work of Jean-Jacques Rousseau, 1712-1754*. University of Chicago Press.

Crawford, K. 2021. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* Yale University Press.

Erdélyi, O. J.; and J. Goldsmith. 2018. Regulating Artificial Intelligence: Proposal for a Global Solution. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA: Association for Computing Machinery. 95–101.

Erman, E.; and M. Furendal. 2024. Artificial Intelligence and the Political Legitimacy of Global Governance. *Political Studies*, 72(2): 421–441.

Estache, A.; G. Garsous; and R. Seroa da Motta. 2016. Shared Mandates, Moral Hazard, and Political (Mis)alignment in a Decentralized Economy. *World Development*, 83: 98–110.

Esteva, A.; B. Kuprel; R. A. Novoa; J. Ko; S. M. Swetter; H. M. Blau; and S. Thrun. 2017. Dermatologist-level Classification of Skin Cancer with Deep Neural Networks. *Nature*, 542(7639): 115–118.

Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

Flathman, R. E. 2002. *Thomas Hobbes: Skepticism, Individuality, and Chastened Politics (Vol. 2)*. Rowman & Littlefield.

Floridi, L. 2014. *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality*. Oxford University Press.

Floridi, L.; J. Cowls; M. Beltrametti; R. Chatila; P. Chazerand; V. Dignum; C. Luetge; R. Madelin; U. Pagallo; F. Rossi; B. Schafer; P. Valcke; and E. Vayena. 2018. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 28(4): 689–707.

Freeman, S. 2007. *Rawls*. London: Routledge.

Frey, C. B.; and M. A. Osborne. 2017. The Future of Employment: How Susceptible are Jobs to Computerisation? *Technological Forecasting and Social Change*, 114: 254–280.

Gabriel, I. 2022. Toward a Theory of Justice for Artificial Intelligence. *Daedalus*, 151(2): 218–231.

Gaus, G. 2010. *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*. Cambridge University Press.

Gauthier, D. 1986. *Morals by Agreement*. Oxford University Press.

Guihot, M.; A. F. Matthew; and N. P. Suzor. 2017. Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence. *Vanderbilt Journal of Entertainment & Technology Law*, 20: 385–456.

Habermas, J. 2015. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. John Wiley & Sons.

Hampton, J. 1986. *Hobbes and the Social Contract Tradition*. Cambridge University Press.

Harris, T. 2007. *Revolution: The Great Crisis of the British Monarchy, 1685-1720*. Penguin UK.

Helbing, D.; B. S. Frey; G. Gigerenzer; E. Hafen; M. Hagner; Y. Hofstetter; J. Van Den Hoven; R. V. Zicari; and A. Zwitter. 2019. Will Democracy Survive Big Data and Artificial Intelligence? In *Towards Digital Enlightenment*, Cham: Springer International Publishing. 73–98.

Hildebrandt, M. 2018. Algorithmic Regulation and the Rule of Law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128): 20170355.

Hiley, D. R. 1990. The Individual and the General Will: Rousseau Reconsidered. *History of Philosophy Quarterly*, 7(2): 159–178.

Hobbes, T. 1651. *Leviathan*. Penguin Books.

Hobbes, T. 1996. *Leviathan, Edited by R. Tuck.* Cambridge: Cambridge University Press.

Hoffmann-Riem, W. 2020. Artificial Intelligence as a Challenge for Law and Regulation. In *Regulating Artificial Intelligence*, Cham: Springer International Publishing. 1–29.

Hoye, J. M. 2017. Obligation and Sovereign Virtue in Hobbes's Leviathan. *The Review of Politics*, 79(1): 23–47.

Hurtgen, J. 1979. Hobbes's Theory of Sovereignty in Leviathan. *Reason Papers*, 5: 55–67.

Kavka, G. S. 2021. *Hobbesian Moral and Political Theory*. Princeton University Press.

Kenyon, J. P. 1986. *The Stuart Constitution, 1603-1688: Documents and Commentary*. Cambridge University Press.

Kleizen, B.; W. Van Dooren; K. Verhoest; and E. Tan. 2023. Do Citizens Trust Trustworthy Artificial Intelligence? Experimental Evidence on the Limits of Ethical AI Measures in Government. *Government Information Quarterly*, 40(4): 101834.

Laux, J. 2024. Institutionalised Distrust and Human Oversight of Artificial Intelligence: Towards a Democratic Design of AI Governance Under the European Union AI Act. *AI & Society*, 39(6): 2853–2866.

Laux, J.; S. Wachter; and B. Mittelstadt. 2024. Trustworthy Artificial Intelligence and the European Union AI Act: On the Conflation of Trustworthiness and Acceptability of Risk. *Regulation & Governance*, 18(1): 3–32.

Leenes, R.; and F. Lucivero. 2014. Laws on Robots, Laws by Robots, Laws in Robots: Regulating Robot Behaviour by Design. *Law, Innovation and Technology*, 6(2): 193–220.

Lin, P.; G. Bekey; and K. Abney. 2008. *Autonomous Military Robotics: Risk, Ethics, and Design*. California Polytechnic State University.

Locke, J. 2011. *Second Treatise of Government*. Hackett Publishing Co.

Lubchenco, J. 1998. Entering the Century of the Environment: A New Social Contract for Science. *Science*, 279(5350): 491–497.

Machado, H.; S. Silva; and L. Neiva. 2023. Publics' Views on Ethical Challenges of Artificial Intelligence: A Scoping Review. *AI and Ethics*.

Martinich, A. P. 2003. *The Two Gods of Leviathan: Thomas Hobbes on Religion and Politics*. Cambridge University Press.

Matthias, A. 2004. The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics and Information Technology*, 6: 175–183.

Mills, C. Wright. 1956. *The Power Elite.* Routledge.

Mills, Charles W. 2019. *The Racial Contract*. Cornell University Press.

Mittelstadt, B. 2019. Principles Alone Cannot Guarantee Ethical AI. *Nature Machine Intelligence*, 1(11): 501–507.

Mittelstadt, B. D.; P. Allo; M. Taddeo; S. Wachter; and L. Floridi. 2016. The Ethics of Algorithms: Mapping the Debate. *Big Data & Society*, 3(2): 2053951716679679.

Molnar, P.; and L. Gill. 2018. Bots at the Gate: A Human Rights Analysis of Automated Decision-Making in Canada's Immigration and Refugee System. *Citizen Lab and International Human Rights Program (Faculty of Law, University of Toronto) Research Report*, 114.

Nozick, R. 1974. *Anarchy, State, and Utopia*. Basic Books.

Okin, S. M. 1989. *Justice, Gender, and the Family*. Basic Books.

O'Neil, C. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Parkin, J. 2007. *Taming the Leviathan: The Reception of the Political and Religious Ideas of Thomas Hobbes in England 1640–1700*. Cambridge University Press.

Pateman, C. 2015. Sexual Contract. In *The Wiley Blackwell Encyclopedia of Gender and Sexuality Studies*, Wiley. 1–3.

Pettit, P. 2012. *On the People's Terms: A Republican Theory and Model of Democracy*. Cambridge University Press.

Pincus, S. C. 2009. *1688: The First Modern Revolution*. Yale University Press.

Power, D. J. 2016. "Big Brother" Can Watch Us. *Journal of Decision Systems*, 25(sup1): 578–588.

Rahwan, I. 2018. Society in the Loop: Programming the Algorithmic Social Contract. *Ethics and Information Technology*, 20(1): 5–14.

Railton, P. 1986. Moral Realism. *The Philosophical Review*, 95(2): 163–207.

Rawls, J. 1971. *A Theory of Justice: Original Edition*. Harvard University Press.

Rawls, J. 1993. *Political Liberalism*. Columbia University Press.

Renda, A. 2019. *Artificial Intelligence. Ethics, Governance and Policy Challenges*. CEPS Centre for European Policy Studies.

Riley, J. 2015. Jean-Jacques Rousseau: The Social Contract. In *Central Works of Philosophy V2*, Routledge. 193–222.

Riley, P. 2013. *Will and Political Legitimacy: A Critical Exposition of Social Contract Theory in Hobbes, Locke, Rousseau, Kant, and Hegel*. Harvard University Press.

Rousseau, J. 1762. The Social Contract (GDH Cole, trans.). *New York: Cosimo Classics*.

Sadler, G. B. 2010. The States of Nature in Hobbes' Leviathan. *Fayetteville State University Government and History Faculty Working Papers*.

Sen, A. 2009. *The Idea of Justice*. Harvard University Press.

Sheehy, B.; and I. Damjanovic. 2023. Social Contract. In *Encyclopedia of Sustainable Management*, Cham: Springer International Publishing. 2983–2986.

Simmons, A. J. 2000. *Justification and Legitimacy: Essays on Rights and Obligations*. Cambridge University Press.

Simmons, A. J. 2014. *On the Edge of Anarchy: Locke, Consent, and the Limits of Society*. Princeton University Press.

Skinner, Q. 1996. *Reason and Rhetoric in the Philosophy of Hobbes*. Cambridge University Press.

Skinner, Q. 2008. *Hobbes and Republican Liberty*. Cambridge University Press.

Smuha, N. A. 2021. From a 'Race to AI' to a 'Race to AI Regulation': Regulatory Competition for Artificial Intelligence. *Law, Innovation and Technology*, 13(1): 57–84.

Snyder Caron, M.; and A. Gupta. 2020. *The Social Contract for AI.* arXiv:2006.08140

Srinivasan, R.; and D. Ghosh. 2023. A New Social Contract for Technology. *Policy & Internet*, 15(1): 117–132.

Taeihagh, A. 2021. Governance of Artificial Intelligence. *Policy and Society*, 40(2): 137–157.

Taeihagh, A.; M. Ramesh; and M. Howlett. 2021. Assessing the Regulatory Challenges of Emerging Disruptive Technologies. *Regulation & Governance*, 15(4): 1009–1019.

Tuckness, A. 2009. *Locke and the Legislative Point of View: Toleration, Contested Principles, and the Law*. Princeton: Princeton University Press.

Turner, I. 2020. Conceptualising a Protection of Liberal Constitutionalism Post 9/11: an Emphasis upon Rights in the Social Contract Philosophy of Thomas Hobbes. *The International Journal of Human Rights*, 24(10): 1475–1498.

Tyler, T. R. 2006. *Why People Obey the Law*. Princeton University Press.

Van Apeldoorn, L. 2020. On the Person and Office of the Sovereign in Hobbes' Leviathan. *British Journal for the History of Philosophy*, 28(1): 49–68.

Wagner, B. 2016. Algorithmic Regulation and the Global Default: Shifting Norms in Internet Technology. *Etikk I Praksis-Nordic Journal of Applied Ethics*, 10(1), 5–13.

Walter, Y. 2024. Managing the Race to the Moon: Global Policy and Governance in Artificial Intelligence Regulation—A Contemporary Overview and an Analysis of Socioeconomic Consequences. *Discover Artificial Intelligence*, 4(1): 14.

Wiegant, D.; A. Dewulf; and J. Van Zeben. 2024. Alignment Mechanisms to Effectively Govern the Sustainable Development Goals. *World Development*, 182: 106721.

Wilson, P. H. 2010. *Europe's Tragedy: A New History of the Thirty Years War*. Penguin UK.

Wirtz, B. W.; J. C. Weyerer; and B. J. Sturm. 2020. The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration. *International Journal of Public Administration*, 43(9): 818–829.

Wolff, J. 2018. *Robert Nozick: Property, Justice and the Minimal State*. John Wiley & Sons.

Yeung, K. 2018. Algorithmic Regulation: A Critical Interrogation. *Regulation & Governance*, 12(4): 505–523.

Zalnieriute, M.; L. B. Moses; and G. Williams. 2019. The Rule of Law and Automation of Government Decision-Making. *The Modern Law Review*, 82(3): 425–455.

Zuboff, S. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs.