DSSR

Discussion Paper No. 111

Mechanism Design with Blockchain Enforcement

Hitoshi Matsushima and Shunya Noda

March, 2020

Data Science and Service Research
Discussion Paper

Center for Data Science and Service Research
Graduate School of Economic and Management
Tohoku University
27-1 Kawauchi, Aobaku
Sendai 980-8576, JAPAN

Mechanism Design with Blockchain Enforcement*

Hitoshi Matsushima† Shunya Noda‡

First Draft: March 14, 2020; Current Draft: March 17, 2020

Abstract

We study the design of self-enforcing mechanisms that rely on neither a trusted third party (e.g., court, trusted mechanism designer) nor a long-term relationship. Instead, we use a smart contract written on blockchains as a commitment device. We design the *digital court*, a smart contract that identifies and punishes agents who reneged on the agreement. The digital court substitutes the role of legal enforcement in the traditional mechanism design paradigm. We show that, any agreement that is implementable with legal enforcement can also be implemented with enforcement by the digital court. To pursue a desirable design of the digital court, we study a way to leverage truthful reports made by a small fraction of behavioral agents. Our digital court has a unique equilibrium as long as there is a positive fraction of behavioral agents, and it gives correct judgment in the equilibrium if honest agents are more likely to exist than dishonest agents. The platform for smart contracts is already ready in 2020; thus, self-enforcing mechanisms proposed in this paper can be used practically, even now. As our digital court can be used for implementing general agreements, it does not leak the detailed information about the agreement even if it is deployed on a public blockchain (e.g., Ethereum) as a smart contract.

JEL Codes: D47, D82, L86

Keywords: Implementation, Decentralized Mechanism, Smart Contract, Oracle Problem, Self-Judgment

---

*This paper supersedes Matsushima (2019), which demonstrated that smart contracts could substitute one of the role of courts but enhance illegal cartelization. This paper further articulates the convenience and social risk of blockchain enforcement by considering a general mechanism design framework and showing the possibility of (i) unique implementation through behavioral mechanism design, (ii) prevention of false charges, and (iii) privacy preservation. This study has been supported by a grant-in-aid for scientific research (KAKENHI, grant numbers: 16H02009, 18K12742) from the Japan Society for the Promotion of Science (JSPS) and the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of the Japanese government as well as by the Center of Advanced Research in Finance at the University of Tokyo and Vancouver School of Economics at the University of British Columbia. We are grateful to Shumpei Goke, Yoshinori Hashimoto, Wei Li, Kyohei Okumura, Daisuke Oyama, Sergei Severinov, Kyungchul (Kevin) Song, and all the participants of the 15th Joint Economics Workshop of the University of Tokyo and Seoul National University (Seoul) for helpful comments. All remaining errors are our own. Hitoshi Matsushima is appointed as a visiting professor at Graduate School of Economics and Management at Tohoku University.

†Department of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. E-mail: hitoshi@e.u-tokyo.ac.jp.

‡Vancouver School of Economics, University of British Columbia, 6000 Iona Dr, Vancouver, BC V6T 1L4, Canada. E-mail: shunya.noda@gmail.com

1 Introduction

In many contracts and mechanisms, parties are often tempted to renege ex post (e.g., buyers may refuse to make payments after the delivery of the good). If each party is afraid of ex post renege in later stages, the parties cannot make a viable agreement (McAfee and McMillan 1987). In the traditional mechanism design paradigm, *legal enforcement* (contract enforcement with court involvement) has been used to deter violations of the agreement. However, legal enforcement is often slow and needs a high judicial cost. To run a mechanism in a decentralized manner, parties can alternatively rely on the *long-term relationship* — if parties can reward honest behaviors with future payoffs, reneging could be prevented (Athey and Segal 2013; Hörner, Takahashi, and Vieille 2015). However, in many applications, agents interact only once.

In this paper, we study a new approach for solving this problem. We construct a *self-enforcing* mechanism that relies on neither legal enforcement nor a long-term relationship. Instead, our mechanism utilizes a *smart contract* (Szabo 1994) deployed on a *blockchain* (Nakamoto 2008) as a commitment device for preventing agents from reneging. The construction of self-enforcing mechanisms with the use of smart contracts is a new mechanism design problem raised by the emergence of blockchains. We design a smart contract named "*digital court*," which substitutes a court in the traditional mechanism paradigm. As in the traditional paradigm, in our framework, all the communications and actions are taken outside of the blockchain (hence, the whole mechanism need not be written as a smart contract). After all the relevant actions are taken, each agent "self-judges" whether each of the other parties followed the agreement, by sending a message to the digital court. Based on agents' reports, a digital court identifies those who reneged and punishes them by executing some automated monetary transfers, leaving no opportunity for reneging again. Foreseeing the punishment in the future, agents find it unprofitable to make any deviation that leaves verifiable evidence. Just like legal enforcement can be used for any purpose, enforcement by digital courts can be used for implementing *any* agreement.

This paper regards the smart contract as a contract written as a computer protocol deployed on a blockchain. A blockchain is a distributed ledger that is managed in a decentralized manner, without relying on trust in any party. *Cryptocurrency* is one of the leading applications of the blockchain technology — by managing the data about the ownership (account balance) of "currencies," a blockchain performs as a peer-to-peer electronic cash system. In contrast to fiat money, cryptocurrencies are programmable. Many blockchains allow users to write a computer protocol that directly accesses the account balance and makes automated monetary (cryptocurrency) transfers. Such computer protocols are called

smart contracts. After the first blockchain, *Bitcoin*, was created, a number of blockchains that can be used as a platform for smart contracts have been launched. Thus, as of 2020, parties are already able to use smart contracts as a tool of mechanism design.

Smart contracts automatically execute contingent transfers in accordance with the pre-agreed protocol and inputs made afterward. As smart contracts can directly update the account balance managed by the blockchain, using a smart contract, users can make a commitment for contingent payments. Taking advantage of this commitment power, untrusted parties may reach a viable agreement without relying on legal enforcement. Previous studies have already proposed various smart contracts for specific applications, e.g., auctions (Galal and Youssef 2019), bilateral trading (Asgaonkar and Krishnamachari 2019), sharing apps (Bogner, Chanson, and Meeuw 2016), and boardroom voting (McCorry, Shahandashti, and Hao 2017). In contrast to them, we study self-enforcement of general abstract mechanisms by designing a smart contract that performs as a court.

Although a smart contract is a useful commitment device for enforcing monetary transfers, it has a number of limitations. Cryptographers, computer scientists, and engineers are actively debating possible technical solutions. However, our approach does not rely on them — we only use the technology that is already incorporated into popular blockchain platforms (such as Ethereum) as of 2020. Hence, all the mechanisms proposed in this paper can be used in the real world right now. Instead of using new cryptographic technologies, our digital-court approach overcomes the following fundamental limitations of smart contracts through mechanism design.

Privacy If we use a public blockchain as a platform, the smart contract deployed on it becomes publicly observable.[1] Hence, if parties write a smart contract intended for a specific purpose, the public can infer the detail of the agreement the parties reached. The versatility of our approach helps parties to keep privacy. Since digital courts can be used for any purpose, the agreement enforced by a digital court cannot be inferred from the structure of the uploaded digital court.

Low Transaction Cost The users of smart contracts must pay commissions to record-keepers every time they input new information and make an operation. As computations

[1]This is because public blockchains allow anyone to work as a record-keeper. Since record-keepers have to check the validity of transaction requests, they must be able to observe the detail of transactions. The cryptography literature has proposed various technical solutions for maintaining privacy (e.g., Zyskind, Nathan, and Pentland 2015; Kosba, Miller, Shi, Wen, and Papamanthou 2016), but our approach does not rely on them.

executed in a smart contract are expensive, the users want to minimize the use of it.[2] Our approach indeed minimizes commissions as we use a smart contract only for punishments. Our result also articulates that punishments are the only part of transactions that must be executed on blockchains.

Payment Only Smart contracts can only enforce transfers of digital assets whose ownership is managed by the relevant blockchain. If all allocations executed by the agreement were completed in a blockchain, parties could make a mechanism self-enforcing by simply writing the mechanism as a smart contract and deploying it on a blockchain. However, as of 2020, besides money (cryptocurrencies), only a very limited type of digital assets is managed by blockchains. Hence, the commitment power of smart contracts is one-sided — although smart contracts can enforce payments, they cannot directly enforce actions that are taken in exchange for payments. This limitation does not matter to our approach. To deter reneging, it suffices to fine deviators.[3]

Finite Message Spaces The size of the data that blockchains can store and process is limited. Hence, a smart contract must not involve infinite message spaces. This technological limitation naturally rules out the use of unbounded mechanisms (e.g., the canonical mechanism of Maskin 1999), which is often criticized in the literature (e.g., Jackson 1992; Abreu and Matsushima 1992). Our construction of digital courts does not take advantage of the unboundedness of the message space.

In our framework, a smart contract is used only for ex post punishments. As such, all communications and actions but judgment are taken place outside of the blockchain. The input of the outside data (who should be punished?) to the smart contract cannot be automated; thus, participants of the mechanism must report the deviators to the digital court

[2] Antonopoulos and Wood (2018) state that *"any computation executed in a smart contract is very expensive and so should be kept as minimal as possible. It is therefore important to identify which aspects of the application need a trusted and decentralized execution platform"* (Chapter 12). Our results indicate that the aspect that needs a trusted execution platform is punishment for deviators. Huberman, Leshno, and Moallemi (2017, 2019); Budish (2018) provide theoretical foundations for the expensiveness of transactions in the long run.

[3] Hypothetically, the ownership of other financial assets, such as fiat money, stocks, and bonds, can also be managed by blockchains. However, as of 2020, there is no blockchain with which parties can write a smart contract that enforces transfers of such assets. Some emerging companies issue digital assets for fundraising, where the ownership of the digital assets is managed by blockchains (*initial coin offering*). Compared with the market of cryptocurrencies, the market size of other digital assets is negligibly small.

Note also that, many (non-digital) assets and services cannot be managed by blockchains even in the future. For example, in a housing market, while the ownership of a house can (hypothetically) be managed by a blockchain, the blockchain cannot directly enforce transfers of possession as it cannot evacuate the residents. Hence, the limited scope problem will continue to exist even after blockchains are widely adopted.

by themselves. As we assume no party is trusted, agents may input incorrect information. If so, the digital court misunderstands the state of the world and cannot execute the intended transfers. This problem is called the *oracle problem* in the blockchain literature. This is the fundamental design question of the digital court.

The oracle problem should be resolved by incentivizing agents to make correct inputs. As in the traditional paradigm, we assume that reneges are verifiable (otherwise, deviations cannot be prevented even when legal enforcement is available). Hence, (i) agents have complete (verified) information about who is "guilty" (i.e., violated the agreement), and (ii) the mechanism should incentivize agents to input truthful information, ideally in the unique equilibrium. This is a new class of *implementation problems* (see Jackson 2001; Maskin and Sjöström 2002 for comprehensive surveys).[4]

In this paper, we tackle three design questions of the digital court: (i) partial implementation, (ii) unique implementation, and (iii) the false charge problem. The false charge problem is a new design problem that has not appeared in the traditional paradigm, whose detail is explained later.

Partial implementation of blockchain enforcement requires to punish agents who violated the agreement in one equilibrium. As we assume that agents are indifferent between whether the other agents are punished or not, this goal is trivial. The digital court decides whether to punish each agent, based on the other agents' message profile. If the sentence is independent of the defendant's message, and each agent obtains no reward from inputting information, then each agent has a weak incentive to make a truthful report to the digital court. Hence, in an equilibrium, the digital court correctly punishes guilty agents. Note also that, even if we strengthen the equilibrium concept to the strict Nash equilibrium, the goal is still trivial — the digital court can incentivize agents to decide the sentence unanimously just by giving penalties when reports are inconsistent (Design 1).

Partial implementation is unsatisfactory. Since agents do not directly care about whether the judgment is fair or not, there are many equilibria in which the digital court misjudges. Ideally, we want to punish guilty agents correctly in any equilibria. However, since we assume that agents' payoffs from the digital court are independent of the guiltiness of the defendant, it is impossible to exclude unwanted equilibria.

To break down the impossibility result, we consider a way to take advantage of the behavioral motivation that is influenced by the guiltiness of defendants. Specifically, we consider intrinsic preferences for conveying honest and dishonest messages to the digital court. We

[4]In the blockchain literature, various systems that correctly input the truthful information about the market has been invented (Peterson, Krug, Zoltu, Williams, and Alexander 2015; Ellis, Juels, and Nazarov 2017; Adler, Berryhill, Veneris, Poulos, Veira, and Kastania 2018). However, no mechanism for helping a small group of agents to input the truthful information about the local information was established in the literature.

keep assuming that most agents are rational and purely motivated by material payoffs. However, we also assume that some agents may be honest and have a psychological incentive to make a truthful report, and other agents may be adversarial and have an incentive to make an untruthful report. Since we focus on an environment in which agents interact only once (i.e., there is no long-term relationship), each agent does not know whether each of the other agents is *truly* honest or not. Hence, each agent has a belief about the other agents' behavioral types. We show that, when agents believe that honest agents are more likely to exist than adversarial agents, there is a mechanism that incentivizes all rational agents to vote for correct judgment.

The design of the proposed digital court (Design 2) is simple. The sentence is determined by a simple majority rule. In addition, similar to Design 1, jurors are punished when their reports are mutually inconsistent. However, Design 2 allows agents to vote not only for 0 (acquittal) or 1 (conviction) but also any fractional value between 0 and 1. Thanks to this feature, when honest agents are more likely to exist than adversarial agents, all the rational agents are incentivized to report a "slightly more honest" message than the other rational agents because rational agents also want to match their message with honest agents'. Accordingly, from the perspective of rational agents, the only reporting strategy that survives iterated elimination of strictly dominated strategies is voting for a correct decision (*unique implementation*). This result is pervasive in the following sense. First, the fraction of honest agents can be arbitrarily small. Second, the mechanism need not identify who is honest. Design 2 performs well as long as agents believe that some of the other agents (may probabilistically) prefer to be honest. Third, the structure of psychological cost functions can be general. Fourth, Design 2 performs well even if there are adversarial agents, as long as their ex ante fraction is smaller than honest agents'.

If we assume that agents have no intrinsic preference, Design 2 fails to implement correct judgment, just as Design 1 fails to do so. However, once we assume the existence of a small fraction of behavioral agents, the equilibria of Design 2 shrinks drastically. Consequently, we can achieve unique implementation of enforcement. How a mechanism leverages the possibility of the existence of behavioral agents is a new criterion for evaluating mechanisms, and our Design 2 makes a difference here.[5]

If adversarial agents are absent, Design 2 makes correct judgment with probability one.

[5]The equilibrium analysis of the game under the presence of (a small fraction of) behavioral agents and incomplete information itself has a long history. For example, Kreps, Milgrom, Roberts, and Wilson (1982) study how the existence of behavioral agents changes the equilibria of finitely repeated games, and Postlewaite and Vives (1987); Carlsson and Van Damme (1993); Morris and Shin (1998) study how incomplete information shrinks the set of equilibria. These previous studies focus on the analysis of given games. In contrast, our focus is on the design of mechanisms that fully take advantage of the existence of behavioral agents and incomplete information.

However, under the presence of adversarial agents, Design 2 may make a *false charge* — if we accidentally have many adversarial agents, the digital court would misunderstand the state of the world. In other words, Design 2 sometimes punishes innocent agents accidentally. Ideally, the mechanism should not fine innocent agents even if some agents prefer to make incorrect inputs. To achieve this goal, we design Design 3, which is a hybrid of Design 1 and Design 2 that incentivizes adversarial agents to tell the truth. Design 3 imposes a large fine when agents' reports are mutually inconsistent, because its payment needs to offset adversarial agents' intrinsic incentives to tell a lie. To avoid fining agents who input truthful information into the digital court, we further elaborate a mechanism. Our final design of the digital court, Design 4, only imposes an arbitrarily small fine to agents who made a truthful input, on the equilibrium path.

On one hand, our results indicate that smart contracts may improve social welfare. Lawsuits and compulsory execution are often costly and time-consuming. If a victim hesitates to pay the judicial cost, lawsuits may become an empty threat and cannot prevent agents from reneging. Even in such situations, punishments by smart contracts could be a credible threat as its cost for enforcement could be significantly lower than that of legal enforcement. Hence, smart contracts may resolve many real-world hold-up problems.

On the other hand, our findings indicate that blockchains and smart contracts may jeopardize the real-world economy more seriously than the previous studies has expected. Unlike real courts, digital courts do not examine the legality of the agreement to be enforced. Therefore, smart contracts enable parties to write viable contracts intended for illegal purposes. For example, Tirole (1992) discusses that side payments can be enforced through a long-term relationship, and therefore, parties can implement an illegal agreement without relying on legal enforcement if agents interact repeatedly. However, our result indicates that under the presence of the smart-contract platform, parties can write a self-enforcing illegal contract, even if their relationship is one-shot. Furthermore, if parties follow our digital-court approach, regulators cannot detect the detail of the agreement by monitoring the smart contract uploaded to a blockchain.[6]

The rest of this paper is organized as follows. Section 2 describes the framework of our decentralized mechanism design problem. Section 3 explains legal enforcement as a benchmark method. Section 4 models smart contracts and digital courts, and explains how they work as a new way to enforce agreements. Section 5 considers partial implementation. Section 6 introduces behavioral aspects of agents' preferences and considers unique implementation.

[6]Although Cong and He (2019) suggest that blockchains may encourage greater collusion, they focus on tacit collusion because *"explicit form of collusion using smart contracts is easy to detect and can be forbidden by antitrust law"* (Cong and He 2019, pp.1729). However, if a cartel follows our approach, the agreement enforced by a digital court cannot be detected from its appearance.

Section 7 presents two applications, auctions and bidding ring, and argues about regulatory policies to prevent the use of digital court for illegal purposes. Section 8 considers the false charge problem. Section 9 concludes.

2 Decentralized Mechanisms

We first introduce a framework of the decentralized mechanism design problem, where the trusted mechanism designer is absent. There are $n \geq 2$ agents $I = \{1, 2, \ldots, n\}$ that attempt to implement a joint decision. Agents are willing to implement an *agreement* $\alpha : \Xi \to A$ that maps a message profile $\xi \in \Xi$ to an action profile $\alpha(\xi) \in A \coloneqq \prod_{i \in I} A_i$. Formally, the game proceeds as follows.

Step 1: A type profile $\theta \coloneqq (\theta_i)_{i \in I} \in \Theta \coloneqq \prod_{i \in I} \Theta_i$ is realized. Every agent i observes her type, $\theta_i \in \Theta_i$.

Step 2: To figure out the state, θ, every agent i publicly announces a message, $\xi_i \in \Xi_i$, simultaneously. The announced message profile $\xi \coloneqq (\xi_i)_{i \in I} \in \Xi \coloneqq \prod_{i \in I} \Xi_i$ becomes public information among agents. Since all agents observe ξ, the agreed action profile $\alpha(\xi)$ becomes common knowledge at this point.

Step 3: Each agent takes an action $\hat{a}_i \in A_i$, simultaneously (simultaneousness is for simplicity and can easily be relaxed). Each agent verifies the action she took to the other agents. Hence, the set of *guilty* agents (deviators), $D(\hat{a}, \xi) \coloneqq \{i \in I : \hat{a}_i \neq \alpha_i(\xi)\}$, becomes common knowledge and verifiable information. We denote agent i's payoff that is finalized in Step 3 by $u_i : A \times \Theta \to \mathbb{R}$. We assume that u_i is bounded, quasi-linear, and risk-neutral.

For notational convenience, we represent the set of deviators by a n-dimensional vector, $\omega \coloneqq (\omega_i)_{i \in I} \in \Omega \coloneqq \{0, 1\}^n$. For each i, ω_i denotes agent i's *guiltiness*: $\omega_i = 0$ if "agent i is *innocent*," i.e., $i \notin D(\hat{a}, \xi)$, and $\omega_i = 1$ if "agent i is *guilty*," i.e., $i \in D(\hat{a}, \xi)$.

Thus far, all the communications and actions are taken outside of the blockchain. Note that, if agents are not willing to use legal enforcement, they need not prepare an explicitly-written contract (or a smart contract) that specifies the agreement to be implemented. See Section 7 for concrete examples of decentralized mechanisms.

If the game terminates in Step 3, the agreement α is just a recommendation. Hence, α is implementable only if agents voluntarily follow the recommendation; i.e., $\alpha(\theta)$ is a (ex post) Nash equilibrium of the normal-form game in which each agent i's utility is specified

by $u_i(\cdot, \theta) : A \to \mathbb{R}$. This is a very stringent condition. For example, the agreement can recommend no monetary transfer because no agent is willing to make a payment voluntarily. This is why mechanisms need enforcement.

To prevent agents from reneging, we want to punish deviations. Let $t_i \in \mathbb{R}_+$ be the fine imposed on agent i ex post. Since we assume that utility functions are quasi-linear, agent i's total payoff is represented in the following form.

$$U_i(\hat{a}, \theta, t_i) = u_i(\hat{a}, \theta) - t_i.$$

We discuss how to implement ex post punishment in detail from the next section.

We define

$$T_i \coloneqq \sup_{\xi \in \Xi, \theta \in \Theta, \hat{a}_i \in A_i} \{u_i(\hat{a}_i, \alpha_{-i}(\xi), \theta) - u_i(\alpha(\xi), \theta)\} + \epsilon,$$

where $\epsilon > 0$ is an arbitrarily small number. When agent i is fined T_i if she is guilty and 0 otherwise, agent i has no incentive for reneging. To enforce the agreement α, it suffices to find a way to fine T_i as a credible threat.

Note that we made three important assumptions on the actions. First, we assume that all the actions $a_i \in A_i$ are taken outside of the blockchain, and therefore, actions cannot be directly enforced by smart contracts. Note that actions may include transfers of fiat money.

Second, we assume that all the monetary transfers that happen on the equilibrium path should be specified as a part of actions because we focus on mechanisms that use transfers in cryptocurrencies only as ex post punishments. This is for keeping privacy: agents do not want to disclose the detailed information about the agreement α, the reported message profile ξ, and taken actions $\hat{a}$ through the smart contract written on a blockchain.

Third, we assume that actions are verifiable. If not, implementation is difficult even when a court is present. The aim of this paper is to discuss how smart contracts may substitute the role of courts in the traditional mechanism design paradigm. We exclude moral hazard problems that are relevant even when legal enforcement is available, by assuming that actions are verifiable.

Remark 1. In Step 2, agents are required to make *public announcements*. Once a public announcement is made, the announced message immediately becomes common knowledge among agents.[7] Hence, once an agent verifies the fact that she followed the agreement, all agents can immediately understand her innocence. Note that public announcements are technically easy. Agents can achieve them by uploading messages to a tamper-proof ledger

[7]Messages need not be disclosed to the outside. To maintain privacy, for example, agents can encrypt all the public announcements by using a common password.

(e.g., a version control system such as GitHub or a blockchain itself).[8]

Remark 2. In Step 2, agents announce messages simultaneously. Even if a trusted mechanism designer is absent and moves are sequential by nature, agents can use a *commitment scheme* to declare messages simultaneously. The commitment scheme is a standard application of cryptography (see, for example, Goldreich 2007).

3 Benchmark: Legal Enforcement

Traditionally, human society has relied on legal enforcement for discouraging parties from reneging. The law prohibits agents from violating the contract, and courts enforce agents to keep the promise by punishing violators. Specifically, agents play the following Step 4 after Step 3.

Step 4: Agents report (sue) the guiltiness vector ω to a court. The court investigates the detail of the agreement and decides whether the defendant should be punished. If agent i is identified to be guilty, she is fined T_i.[9]

Towards the implementation of the agreement, courts have two important roles. First, a court must investigate the validity of agents' claim, and judge whether agents who are claimed to be guilty should be punished. Recall that the action profile is assumed to be verifiable. Hence, agents cannot misreport the guiltiness to the court. Therefore, this procedure is mathematically easy, as long as the court is non-strategic. For Step 4 to be a credible threat, the court must be *trusted* in the sense that agents believe that the court will punish guilty agents properly.

Second, a court must check whether the agreement is legal and help parties to implement the agreement only when the agreement is made for a good purpose. This is for protecting society from the threat of illegal activities. Thanks to this function, criminals have not been able to use a mechanism that relies on legal enforcement.

[8] Akbarpour and Li (2019) study an auction design problem in which legal enforcement is available but a trusted mechanism designer is absent. Their environment is crucially different from ours in that they exclude public announcements by assumption (in their model, only private communications between each agent and an untrusted mechanism designer are allowed). The middleman problem they study does not appear if public announcements are available.

[9] In reality, when the court admits a violation of the contract, the court typically orders the defendant to pay compensation to the plaintiff. From the mechanism design perspective, both compensation and fine work as punishments for deterring contract violations. Hence, we do not strictly distinguish these two terminologies.

Even if courts are trusted, legal enforcement is not almighty. First, lawsuits are costly. The capacity of courts is always limited, and decision making is typically slow. If ex post punishment becomes an empty threat for these reasons, then deviations cannot be prevented. Second, under some circumstances, agents may not want to rely on courts. For example, agents might be concerned about privacy. If they rely on a lawsuit, the relevant information will be disclosed to the public.

4 Blockchain Enforcement

4.1 Smart Contract

First, we model smart contracts mathematically.

Definition 1 (Smart Contract)**.** A *smart contract* $(M, \bar{t}, \gamma)$ is a triple of a message space $M \coloneqq (M_i)_{i \in I}$, a deposit vector $\bar{t} \coloneqq (\bar{t}_i)_{i \in I} \in \mathbb{R}^n_+$, and a transfer rule $\gamma : M \to \prod_{i \in I}(-\infty, \bar{t}_i]$, where the transfer rule γ is weakly budget balanced; i.e., $\sum_{i \in I} \gamma_i(m) \geq 0$ for all $m \in M$.

The smart contract initially receives $\bar{t}_i$ from agent i as a deposit. When message profile m is input, the smart contract (re)pays $\bar{t}_i - \gamma_i(m)$ to agent i. Hence, agent i's net payment is $\gamma_i(m)$. The rest of the deposit, $\sum_{i \in I} \gamma_i(m)$, is burned. Once a smart contract is deployed on a blockchain and agents make a deposit, no agent can renege on it.

Here, we define the smart contract as a commitment device for input-contingent payments. This assumption is realistic. First, a smart contract can directly update the account balances of participants, and no one can tamper the smart contract. Thus, through a smart contract, agents can make a commitment to contingent cryptocurrency transfers. Second, as of 2020, besides cryptocurrencies, there are few digital assets whose ownership is managed by blockchains. Hence, the smart contract can hardly enforce transfers other than cryptocurrencies. Third, cryptocurrencies are actively traded on the market as a kind of liquid speculative assets; thus, it is easy to exchange popular cryptocurrencies (such as Bitcoin and Ethereum) for fiat money.[10] Hence, transfers in cryptocurrencies are equivalent to transfers in fiat money.

Unlike courts, smart contracts cannot force agents to pay a fine unless they agree to do so ex ante. Hence, before exchanging information about θ but after making an agreement for the action profile, each agent i needs to deposit $\bar{t}_i = \max_{m \in M} \gamma_i(m)$ to the smart contract.

[10]The exchange market of emerging cryptocurrencies is thin. Hence, such cryptocurrencies might be illiquid and may not be regarded as equivalent to fiat money. This is the reason why we develop a digital court based on the technology widely adopted by popular blockchains.

Step 1+: A type profile $\theta = (\theta_i)_{i \in I}$ is realized. Every agent i observes her type, $\theta_i \in \Theta_i$. At this moment, agents deploy a smart contract $(M, \bar{t}, \gamma)$ on a blockchain. Every agent i deposits $\bar{t}_i$ to the smart contract.

Remark 3. Unless agents can make a deposit, the smart contract cannot enforce cryptocurrency transfer. If agents have a severe liquidity constraint, the digital-court approach might not be available.

There is no change in Step 2 and 3. In Step 4+, agents send the information about the set of guilty agents, ω (or $D(\hat{a}, \xi)$), to the smart contract as a message m_i, rather than a court. The smart contract should identify guilty agents from the input (message profile) and punish agents who are identified to be guilty.

Step 4+: Every agent i sends a message m_i to the smart contract. The smart contract returns $\bar{t}_i - \gamma_i(m)$ to agent i. After the repayment, the smart contract is cleared.

Beware that our model involves two different messages. Messages ξ is communication between agents, that is used for deciding the action profile that the agents agreed to take, $\alpha(\xi)$. Messages m is an input into the smart contract, $(M, \bar{t}, \gamma)$. Each agent i's message space M_i is explicitly defined in the smart contract, and agent i submits a message m_i as a transaction on the blockchain. Messages m is used for inputting information about the guiltiness of agents, ω, and based on this message profile, the smart contract executes automated monetary transfers, $\gamma(m)$.

Agent i's resultant payoff is

$$u_i(\hat{a}, \theta) - \gamma_i(m).$$

In Step 4+, the message profile ξ is already reported, and the action profile $\hat{a}$ is already taken. Accordingly, they do not directly influence the incentives for messaging. Hence, in any perfect Bayesian equilibrium, in Step 4+, each agent i chooses a message that minimizes her fine $\gamma_i(m)$; i.e.,

$$\gamma_i(m_i, m_{-i}) \leq \gamma_i(m_i', m_{-i}) \text{ for all } m_i' \in M_i. \tag{1}$$

We say that a message profile m is a *Nash equilibrium* of a smart contract $(M, \bar{t}, \gamma)$ if it satisfies (1) for every $i \in I$. We say that the message profile m is a *strict Nash equilibrium* of a smart contract $(M, \bar{t}, \gamma)$ if (1) is satisfied with strict inequalities for all $i \in I$ and $m_i' \in M_i \setminus \{m_i\}$.

4.2 Digital Court

Although smart contracts can be used for various purposes, we only use it for preventing agents from violating the agreement, i.e., to take $\hat{a}_i \neq \alpha_i(\xi)$. To achieve this goal, it suffices to focus on the design of a simple class of smart contracts, named *digital courts*.

Definition 2. A smart contract $(M, \bar{t}, \gamma)$ is called a *digital court* if the following conditions are satisfied.

- For every $j \in I$, $M_j \coloneqq \prod_{i \in I} M_j^i$, where $M_j^i \subseteq [0,1]^K$ for some $K \in \mathbb{Z}_+$.
- For every $j \in I$, γ_j can be represented as

$$\gamma_j(m) = T_j \cdot s^j(m_{-j}^j) + \sum_{i \in I} q_j^i(m^i),$$

  where $s^j : M_{-j}^j \to \{0,1\}$ is called the *sentence function* for agent j, and $q_j^i : M^i \to \mathbb{R}_+$ is the *incentive payment term*. Here, $M_{-j}^j \coloneqq \prod_{i \neq j} M_i^j$ and $M^i \coloneqq \prod_{k \in I} M_k^i$.

Note that a digital court never "rewards" agents in the sense that $\gamma_j(m) \geq 0$ holds for all $j \in I$ and $m \in M$.

A digital court comprises n independent trials, indexed by $i \in I$. Each trial i regards agent i as a *defendant* and determines whether to convict agent i, and all agents (including the defendant herself) participate as *jurors*. Towards each trial i, each agent $j \in I$ expresses her opinion about defendant i's guiltiness, ω_i, by sending a message $m_j^i \in M_j^i$. The profile of messages submitted to each trial, $m_j \coloneqq (m_j^i)_{i \in I}$, is the whole message that agent j sends to the digital court.

Agent i's *sentence function* $s^i : M_{-i}^i \to \{0,1\}$ decides whether to convict agent i based on all the other agents' report for trial i, m_{-i}^i. If agent i is convicted (i.e., $s^i(m_{-i}^i) = 1$), then she is fined T_i. The design objective of the digital court is to convict guilty agents and acquit innocent agents; i.e., to achieve $s^i(m_{-i}^i) = \omega_i$ in an equilibrium. To decide the sentence, the digital court only looks at the messages sent from agents other than agent i herself (otherwise, agent i is strongly incentivized to insist on her innocence in trial i).

In this paper, we focus on digital courts whose message space is a subset of a (multi-dimensional) unit interval. Sending a message closer to 0 means that the juror votes to acquit the defendant, and sending a message closer to 1 means that the juror votes to convict the defendant. All the sentence functions we consider in this paper is monotonic in the sense that the defendant is more likely to be convicted (i.e., s^i becomes more likely to take 1) if each juror $j \in I$ reports a larger message to trial i.

To incentivize each agent to send an intended message, the digital court also involves an *incentive payment term*, $q_j^i : M^i \to \mathbb{R}_+$, which is a fine imposed on juror $j \in I$ for her activity in trial $i \in I$. If the message profile sent to trial i is m^i, juror j is fined $q_j^i(m^i)$. Since the sentence function s^j is independent of agent j's own message m_j^j, agent j's incentive for reporting is solely provided by $\sum_{i \in I} q_j^i(m^i)$. Furthermore, since each incentive payment term for trial i only takes account of the message profile sent to trial i, to minimize $\sum_{i \in I} q_j^i(m^i)$, it suffices to minimize $q_j^i(m^i)$ separately, for each i.

5 Partial Implementation

5.1 Weak Incentives

We start from discussing partial implementation of the punishment; i.e., to achieve $s^i(m_{-i}^i) = \omega_i$ in a Nash equilibrium.

This goal is trivial. Let $M_j^i = \{0, 1\}$ for all $i, j \in I$. Each agent j is expected to match her report with the defendant's guiltiness, i.e., send $m_j^i = \omega_i$. If the incentive payment is set to zero for all message profiles; i.e., $q_j^i(m^i) = 0$ for all $m^i \in M^i$, then agent j is indifferent between sending any $m_j^i \in M_j^i$. Accordingly, telling the truth is one of the best responses. In a trivial Nash equilibrium, every agent j sends $m_j^i = \omega_i$. For example, we can decide the sentence by a majority rule:

$$s^i(m_{-i}^i) = \begin{cases} 1 & \text{if } \displaystyle\sum_{j \neq i} m_j^i > \frac{n-1}{2}, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Then, in an equilibrium, every agent sends $m_j^i = \omega_i$. Therefore, $s^i(m_{-i}^i) = \omega_i$ is also achieved in a Nash equilibrium.

Note that, since decisions are made unanimously in a Nash equilibrium, we need not use a simple majority rule as a sentence function.

5.2 Strict Incentives

In many applications, implementations that rely on weak incentives are not trustworthy. It is desirable to provide strict incentives for taking intended actions. This is not difficult either. We can incentivize jurors to make a decision unanimously by comparing multiple jurors' messages.

Definition 3. *Design 1* of the digital court is specified as follows:

- $M_j^i \coloneqq \{0,1\}$ for all $i,j \in I$.
- Defendant i's sentence function s^i is a simple majority rule as specified by (2).
- Juror j's incentive payment term for trial i, q_j^i, is given by

$$q_j^i(m^i) \coloneqq \eta \cdot \mathbb{1}\left\{m_j^i \neq m_k^i \text{ for some } j,k \in I\right\},$$

  where $\eta > 0$.
- Agent j's deposit $\bar{t}_j$ is given by

$$\bar{t}_j \coloneqq T_j + n \cdot \eta.$$

If votes for trial i is unanimous, i.e., $m_j^i = m_k^i$ for all $j,k \in I$, then the digital court repays $\eta > 0$ to every juror $j \in I$. Otherwise, the deposit η is not returned to any agent $j \in I$, and it is burned instead. Since agent j participates in n different trials as a juror, if her choice matches the aggregate decision for all trials, then she can be repaid with $n\eta$ in total.

Design 1 encourages jurors to agree on the sentence unanimously. In one equilibrium, jurors match their votes at the correct reporting.

Theorem 1. *Design 1 has a strict Nash equilibrium in which every agent makes truthful reporting to the digital court; i.e., every agent j reports $m_j = \omega$. In this equilibrium, $s^i(m^i_{-i}) = \omega_i$ for all $i \in I$.*

Proof. The incentive payment q_j^i rewards agent j only if her message matches all the other jurors $I \setminus \{j\}$. Hence, for every agent $j \in I$, given that all the other agents $k \in I \setminus \{j\}$ reports $m_k^i = \omega_i$, agent j's unique best response is to report $m_j^i = \omega_i$. Hence, $m_j = \omega$ comprises a strict Nash equilibrium. □

6 Unique Implementation

6.1 Impossibility Result

Design 1 implements the agreement as one equilibrium, but it is not unique implementation. Design 1 indeed has many Nash equilibria. While Design 1 incentivizes jurors to match their messages, the focal point need not be truthtelling. For example, a strategy profile in which every agent $j \in I$ tells a lie; i.e., $m_j^i = 1 - \omega_i$ for all $i \in I$, also constitutes a Nash equilibrium.

It would be ideal if we could achieve unique implementation of correct judgment. However, in Step 4+, an action profile $\hat{a}$ and type profile θ are already finalized. Hence, $u_i(\hat{a}, \theta)$ is constant, and agents' payoffs are solely determined by monetary transfers through the digital court. Since agents' preferences for monetary transfers are independent of types (guiltiness), unique implementation is impossible.

Fact. *For any smart contract $(M, \bar{t}, \gamma)$, the set of Nash equilibria does not depend on ω.*

Proof. Each agent's payoff function is independent of ω. Hence, the set of Nash equilibria cannot depend on ω. □

Note that, even if we introduce a refined solution concept, such as subgame perfect implementation (Moore and Repullo 1988), Nash implementation with undominated strategies (Palfrey and Srivastava 1991), and weak iterative dominance (Abreu and Matsushima 1994), the impossibility results cannot be broken down. Moreover, (not exact but) virtual implementation (Matsushima 1988; Abreu and Sen 1991; Abreu and Matsushima 1992) is also impossible. This is because mechanisms that have been developed in the standard implementation theory literature crucially rely on the assumption that agents' preferences are influenced by the state of the world. However, in our problem, the state (guiltiness) does not change agents' preferences over monetary transfers.

6.2 Intrinsic Preferences

To break down the impossibility result above, we must leverage behavioral aspects of economic agents. Although the mechanism design literature has typically assumed that agents are purely interested in material payoffs, experimental research has suggested that some people have intrinsic preferences for honest behaviors (Gneezy 2005; Abeler, Nosenzo, and Raymond 2019). We assume that, with a small probability, there are honest agents who prefer to convey truthful information to the digital court. At the same time, we also assume that some agents might be adversarial and prefer to tell a lie.

Let $B_j := \{R, H, A\}$ be agent j's *behavioral type space*. $b_j = R$ means "agent j is *rational*," $b_j = H$ means "agent j is *honest*," and $b_j = A$ means "agent j is *adversarial*." From now, we take account of each behavioral type of agents' psychological cost from reporting, and we assume that each agent j minimizes her expected *disutility*, $\Gamma_j : M \times \Omega \times B_j \to \mathbb{R}$, which is the sum of her fine and psychological cost.

First, we consider the following form of psychological costs.

Assumption 1. For every $i, j \in I$, $M_j^i \subseteq [0,1]$. For every agent $j \in I$, with probability $1 - \delta_H - \delta_A$, agent j is *rational* ($b_j = R$) and wants to minimize

$$\Gamma_j(m, \omega, R) \coloneqq \gamma_j(m).$$

With probability $\delta_H \geq 0$, agent j is *honest* ($b_j = H$) and wants to minimize

$$\Gamma_j(m, \omega, H) \coloneqq \gamma_j(m) + \sum_{i \in I} \left[\omega_i \cdot c_H^1(m_j^i) + (1 - \omega_i) \cdot c_H^0(m_j^i)\right],$$

where (i) both $c_H^1 : [0,1] \to \mathbb{R}_+$ and $c_H^0 : [0,1] \to \mathbb{R}_+$ are strictly convex and twice differentiable, (ii) c_H^1 is strictly decreasing and $c_H^1(1) = 0$, and (iii) c_H^0 is strictly increasing and $c_H^0(0) = 0$. With probability $\delta_A \geq 0$, agent j is *adversarial* ($b_j = A$) and wants to minimize:

$$\Gamma_j(m, \omega, A) \coloneqq \gamma_j(m) + \sum_{i \in I} \left[\omega_i \cdot c_A^1(m_j^i) + (1 - \omega_i) \cdot c_A^0(m_j^i)\right],$$

where (i) both $c_A^1 : [0,1] \to \mathbb{R}_+$ and $c_A^0 : [0,1] \to \mathbb{R}_+$ are strictly convex and twice differentiable, (ii) c_A^1 is strictly increasing and $c_A^1(0) = 0$, and (iii) c_A^0 is strictly decreasing and $c_A^0(1) = 0$.

The realization for each agent is independent, and each agent cannot observe whether other agent is honest or not.

Later, we will consider a mechanism (Design 2) that allows agents to cast a fractional vote, $m_j^i \in [0,1]$. Thus, the domain of the psychological cost functions is also extended from $\{0,1\}$ to $[0,1]$.

Note that, Assumption 1 does not assume the (ex post) existence of behavioral agents. When δ_H and δ_A are small, it is likely that all parties are rational and want to minimize the fine. Note also that we assume that agents' behavioral types are not observable from the outside, and therefore, the mechanism cannot use a behavioral agent's opinion as a reference.[11]

For example, Assumption 1 is satisfied if the psychological cost functions are quadratic: $c_H^1(m_j^i) = c_A^0(m_j^i) = (1 - m_j^i)^2$ and $c_H^0(m_j^i) = c_A^1(m_j^i) = (m_j^i)^2$. Such a case is investigated in more detail in Subsection 6.6.

The first term, $\gamma_j(m)$, is the total fine imposed on agent j, and rational agents are interested only in it. If agent j is either honest or adversarial, she also takes account of

[11] Matsushima (2008) investigates unique implementation of social choice functions in an environment where there exists a player who has intrinsic preference for honesty and whose opinions can be treated as a reference.

psychological costs. For example, if defendant i is guilty (i.e., $\omega_i = 1$), an honest juror j incurs a psychological cost $c_H^1(m_j^i)$ by reporting m_j^i. The psychological cost is zero if agent j tells the truth ($m_j^i = 1$ leads to $c_H^1(1) = 0$) and becomes larger as the message becomes less truthful (i.e., $c_H^1(m_j^i)$ is strictly decreasing). The payoff structure of adversarial agents is similar, but an adversarial agent incurs a larger psychological cost as she sends a more truthful message.

As we have introduced behavioral types and incomplete information into the model, we should also define Bayesian Nash equilibrium of digital courts. Let $\sigma_j : \Omega \times B_j \to M_j$ be agent j's *reporting strategy*, and $\sigma \coloneqq (\sigma_j)_{j \in I}$ be the *reporting strategy profile*.

Definition 4. We say that a reporting strategy profile σ is a *Bayesian Nash equilibrium* of a smart contract $(M, \bar{t}, \gamma)$ if it satisfies

$$\mathbb{E}\left[\Gamma_j\left(\sigma_j(\omega, b_j), \sigma_{-j}(\omega, \tilde{b}_{-j}), b_j\right)\right] \leq \mathbb{E}\left[\Gamma_j\left(m'_j, \sigma_{-j}(\omega, \tilde{b}_{-j}), b_j\right)\right]$$

for all $j \in I$, $\omega \in \Omega$, $b_j \in B_j$.

6.3 Design 1 Still Fails

Assumption 1 does not lead a naïve design of the digital court to an ideal outcome. As an optimistic scenario, let us assume that there is no adversarial agent who intrinsically prefers to tell a lie (i.e., $\delta_A = 0$) but there is a large fraction of honest agents (i.e., $\delta_H \gg 0$).

In the best case, all honest agents would be motivated by their intrinsic preferences and report $m_j^i = \omega_i$ to the digital court. The existence of honest agents encourages rational agents to tell the truth because rational agents want to match their messages with the other agents'. However, even when we assume Assumption 1 with $\delta_H > 0$ and $\delta_A = 0$, Design 1 may still have multiple equilibria with a wide range of δ_H. If $\delta_H < 1/2$ and a rational agent believes that all the other rational agents make an untruthful report, a decision is more likely to be made unanimously if the agent also report untruthfully. Hence, in an equilibrium, all rational agents may still tell a lie. The equilibrium becomes unique if $\delta_H > 1/2$. However, in many applications, we expect that most of economic agents are rational, and therefore, $\delta_H > 1/2$ is a too stringent assumption.

6.4 Unique Equilibrium

We construct an alternative digital court, Design 2, that has a unique equilibrium under a generic range of parameters. Design 2 is defined as follows:

Definition 5. *Design 2* of the digital court is specified as follows:

- $M_j^i = [0,1]$ for all $i, j \in I$.
- Defendant i's sentence function s^i is given by

$$s^i(m_{-i}^i) \coloneqq \begin{cases} 1 & \text{if } \sum_{j \neq i} \mathbb{1}\left\{m_j^i > \frac{1}{2}\right\} > \frac{n-1}{2}, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

- Juror j's incentive payment term for trial i, q_j^i, is given by

$$q_j^i(m^i) \coloneqq \frac{\eta}{n-1} \cdot \sum_{k \neq j} (m_j^i - m_k^i)^2.$$

  where $\eta > 0$.

- Agent j's deposit $\bar{t}_j$ is given by

$$\bar{t}_j \coloneqq T_j + \eta \cdot n.$$

Unlike Design 1, Design 2 allows jurors to cast a fractional vote. For each trial i, agent $j \in I$ reports a fractional value $m_j^i \in [0,1]$, where $m_j^i = 0$ means "agent i is surely innocent" and $m_j^i = 1$ means "agent i is surely guilty."

Remark 4. To be precise, a continuous message space (like $[0,1]$) cannot be used in smart contracts because the data size of inputs must be finite. Hence, the message space of Design 2 must be approximated by a fine-grained discrete message space.

Similar to Design 1, the sentence function of Design 2, (3), is based on a majority rule. Here, messaging $m_j^i > 1/2$ means voting for "guilty," and $m_j^i \leq 1/2$ means voting for "innocent." The digital court counts the number of votes for "guilty" and imposes a fine to the defendant if a majority of jurors made such votes. We say that juror j *is voting for a correct decision* if either (i) $m_j^i > 1/2$ and $\omega_i = 1$ or (ii) $m_j^i < 1/2$ and $\omega_i = 0$. If a defendant i is convicted, she loses T_i.

Each juror j's fine for trial i is calculated by comparing j's message m_j^i and another juror k's message m_k^i, one by one. Each incentive payment term is specified by $(m_j^i - m_k^i)^2$. This term incentivizes agent j to report her best guess of the opponent k's message — to minimize the fine term, agent j should choose $m_j^i = m_k^i$, and then agent j is not fined. Conversely, the fine is maximized when agent j and k completely disagree, i.e., either (i) $m_j^i = 0$ and $m_k^i = 1$ or (ii) $m_j^i = 1$ and $m_k^i = 0$. In such cases, $(m_j^i - m_k^i)^2 = 1$.

Parallel to Design 1, Design 2 incentivizes rational agents to match their messages with the others. However, in contrast to Design 1, Design 2 allows agents to report a fractional value. Since behavioral agents also have intrinsic preferences, they have an incentive to make their reports slightly closer to their bliss point. Since rational agents also expect this, as we proceed the iterated elimination of strictly dominated strategies, the set of rationalizable strategies gradually shrinks. Eventually, we reach a unique strategy profile.

Theorem 2. *Suppose Assumption 1 and either $\delta_H > 0$ or $\delta_A > 0$. Then, Design 2 is dominance solvable and has a unique Bayesian Nash equilibrium.*

Proof. Since each trial is independent and the mechanism is symmetric, we only consider an innocent agent i's trial ($\omega_i = 0$). We prove the uniqueness of the equilibrium by performing iterated strict dominance. Let $M_R(0) = M_H(0) = M_A(0) = [0, 1]$ and $M_R(r), M_H(r), M_A(r)$ be the set of undominated strategies in r-th round from the perspective of rational, honest, and adversarial agents, respectively.

By choosing m_j^i, a rational agent j minimizes $q_j^i(m^i)$. An honest agent j minimizes $q_j^i(m^i) + c_H^0(m_j^i)$. An adversarial agent j minimizes $q_j^i(m^i) + c_A^0(m_j^i)$. Since the psychological cost functions are independent of the other agents' reports, the strategic complementarity of this game is solely provided by the incentive payment term, $q_j^i(m^i)$. Clearly, the fine is imposed according to a submodular function, and therefore, the game implied by Design 2 is a supermodular game. Hence, to proceed the iterated elimination of strictly dominated strategies, it suffices to consider the behaviors of $\max M_R(r)$, $\max M_H(r)$, $\max M_A(r)$, $\min M_R(r)$, $\min M_H(r)$, $\min M_A(r)$. They are given by

$$
\begin{aligned}
\max M_R(r+1) &= \underset{m_j^i \in M_R(r)}{\arg\min}\ \eta \cdot \bar{Q}(m_j^i; r), \\
\max M_H(r+1) &= \underset{m_j^i \in M_H(r)}{\arg\min}\ \left\{\eta \cdot \bar{Q}(m_j^i; r) + c_H^0(m_j^i)\right\}, \\
\max M_A(r+1) &= \underset{m_j^i \in M_A(r)}{\arg\min}\ \left\{\eta \cdot \bar{Q}(m_j^i; r) + c_A^0(m_j^i)\right\}, \\
\min M_R(r+1) &= \underset{m_j^i \in M_R(r)}{\arg\min}\ \eta \cdot \underline{Q}(m_j^i; r), \\
\min M_H(r+1) &= \underset{m_j^i \in M_H(r)}{\arg\min}\ \left\{\eta \cdot \underline{Q}(m_j^i; r) + c_H^0(m_j^i)\right\}, \\
\min M_A(r+1) &= \underset{m_j^i \in M_A(r)}{\arg\min}\ \left\{\eta \cdot \underline{Q}(m_j^i; r) + c_A^0(m_j^i)\right\},
\end{aligned}
$$

where

$$\begin{aligned}\bar{Q}(m_j^i; r) &= (1-\delta_H-\delta_A)\left(m_j^i - \max M_R(r)\right)^2 \\ &\quad + \delta_H\left(m_j^i - \max M_H(r)\right)^2 + \delta_A\left(m_j^i - \max M_A(r)\right)^2, \\ \underline{Q}(m_j^i; r) &= (1-\delta_H-\delta_A)\left(m_j^i - \min M_R(r)\right)^2 \\ &\quad + \delta_H\left(m_j^i - \min M_H(r)\right)^2 + \delta_A\left(m_j^i - \min M_A(r)\right)^2.\end{aligned}$$

Since the set of messages that survive iterated strict dominance is monotonic, its maximum and minimum value, $\max M_R(r), \min M_R(r)$, etc., are convergent sequences. Therefore, if M_R^*, M_H^*, M_A^* are the limit of either $\max M_R(r)$, $\max M_H(r)$, $\max M_A(r)$ or $\min M_R(r)$, $\min M_H(r)$, $\min M_A(r)$, then M_R^*, M_H^*, M_A^* must satisfy

$$\begin{aligned}M_R^* &= \arg\min_{m_j^i \in [0,1]} \eta \cdot Q^*(m_j^i), \\ M_H^* &= \arg\min_{m_j^i \in [0,1]} \left\{\eta \cdot Q^*(m_j^i) + c_H^0(m_j^i)\right\}, \\ M_A^* &= \arg\min_{m_j^i \in [0,1]} \left\{\eta \cdot Q^*(m_j^i) + c_A^0(m_j^i)\right\},\end{aligned}$$

where

$$Q^*(m_j^i) = (1-\delta_H-\delta_A)\left(m_j^i - M_R^*\right)^2 + \delta_H\left(m_j^i - M_H^*\right)^2 + \delta_A\left(m_j^i - M_A^*\right)^2.$$

Solving the optimization problems, we have

$$M_R^* = \frac{\delta_H}{\delta_H+\delta_A} M_H^* + \frac{\delta_A}{\delta_H+\delta_A} M_A^*, \tag{4}$$

$$M_R^* = M_H^* + \frac{1}{2\eta}\left(c_H^0\right)'(M_H^*), \tag{5}$$

$$M_R^* = M_A^* + \frac{1}{2\eta}\left(c_A^0\right)'(M_A^*). \tag{6}$$

Note that $\delta_H + \delta_A > 0$ because we assume either $\delta_H > 0$ or $\delta_A > 0$. Equations (5) and (6) regard M_R^* as a function of M_H^* and M_A^*, respectively. Since $(c_H^0)'$ and $(c_A^0)'$ are strictly increasing, the whole functions are also strictly increasing, and therefore, we can take an inverse function. If we regard M_H^* and M_A^* as functions of M_R^*, then it follows from the

inverse function theorem that

$$(M_H^*)'(M_R^*) = \left[1 + \frac{1}{2\eta}\left(c_H^0\right)''\left(M_H^*(M_R^*)\right)\right]^{-1},$$
$$(M_A^*)'(M_R^*) = \left[1 + \frac{1}{2\eta}\left(c_A^0\right)''\left(M_A^*(M_R^*)\right)\right]^{-1}.$$

Since $(c_H^0)''$ and $(c_H^0)''$ are positive, $0 < (M_H^*)'(M_R^*) < 1$ and $0 < (M_A^*)'(M_R^*) < 1$ holds for all $M_R^* \in [0,1]$. Hence, there exists unique $M_R^* \in [0,1]$ that satisfies the fixed point indicated by (4); i.e.,

$$M_R^* = \frac{\delta_H}{\delta_H + \delta_A} M_H^*(M_R^*) + \frac{\delta_A}{\delta_H + \delta_A} M_A^*(M_R^*).$$

Accordingly, iterated strict dominance leads us to a unique Bayesian Nash equilibrium. □

If we assume that all agents are rational and purely interested in material payoffs, then Design 2 has many equilibria, just as Design 1 does. However, once we assume the existence of behavioral agents, the equilibria under these two digital courts are drastically different. While Design 1 remains to have many equilibria under a wide range of parameters, Design 2 always has a unique Bayesian Nash equilibrium as long as behavioral agents exist with a positive probability.

6.5 Unique Implementation of the Judgment

Theorem 2 assures that as long as there is a small chance to have behavioral agents, Design 2 has a unique Bayesian Nash equilibrium. However, Theorem 2 remains silent as to the sentence Design 2 is likely to give. Indeed, the messages submitted in the unique equilibrium are difficult to characterize, as they not only depend on the parameter values of δ_H and δ_A but also the functional form of c_H^0, c_H^1, c_A^0, and c_A^1.

Although characterization of equilibria with general parameters is difficult, it is relatively easy to characterize the equilibrium messages when the scale of incentive payments is small. As long as psychological cost functions are strictly monotonic, at the limit of $\eta \to 0$, all behavioral agents ignore material payoffs and minimize their psychological costs — all honest agents tell the truth and all adversarial agents tell a lie. Hence, rational agents can infer that δ_H fraction of agents report ω_i and δ_A fraction of agents report $1-\omega_i$. Using this fact, we can derive the equilibrium messages of rational agents in a closed form, as a function of ω_i, δ_H and δ_A. The following theorem indicates that, regardless of the shape of psychological cost functions, Design 2 with small η can induce desirable voting of rational agents if $\delta_H > \delta_A$.

Theorem 3. *Suppose Assumption 1 and either $\delta_H > 0$ or $\delta_A > 0$. Then, in the limit of $\eta \to$*

0, *the following reporting strategy profile constitutes a unique Bayesian Nash equilibrium:*

$$\begin{aligned}\sigma_j^i(\omega, H) &= \omega_i, \\ \sigma_j^i(\omega, A) &= 1 - \omega_i, \\ \sigma_j^i(\omega, R) &= \omega_i \cdot \frac{\delta_H}{\delta_H + \delta_A} + (1 - \omega_i) \cdot \frac{\delta_A}{\delta_H + \delta_A}.\end{aligned}$$

Hence, rational agents vote for a correct decision if $\delta_H > \delta_A$.

Proof. Again, we focus on trial i and the case in which defendant i is innocent. Since the unique Bayesian Nash equilibrium of Design 2 is characterized by (4), (5), and (6), it suffices to consider the behavior of M_R^*, M_H^*, and M_A^* in the limit of $\eta \to 0$. Since $(c_H^0)' > 0$ and $(c_A^0)' < 0$ in $(0, 1)$, in the limit of $\eta \to 0$, $M_H^* \to 0$ and $M_A^* \to 1$ must hold. Therefore, (4) implies $M_R^* = \delta_A/(\delta_H + \delta_A)$, as desired. □

If η is large, the messages sent by rational agents depend on the shape of psychological cost functions. For example, even if δ_H is large and δ_A is small, if adversarial agents are extremely stubborn and incurs a large psychological cost if they compromise a little bit, honest agents may give up and make less truthful reports so as to increase their material payoffs. Hence, unless we consider the limit of $\eta \to 0$, the shape of cost functions, c_H^1, c_H^0, c_A^1, and c_A^0, crucially influence the equilibrium messages. Once we assume $\eta \to 0$, the shape of cost functions no longer matters. (See also Subsection 6.6.)

In many applications, we assume that most agents are rational. Hence, we want to make rational agents to vote for a correct decision. In Design 2 with $\eta \to 0$, this condition is satisfied if and only if honest agents are more likely to exist than adversarial agents; i.e., $\delta_H > \delta_A$. In such a case, Design 2 is able to give a correct sentence with a large probability.

If $\delta_H > \delta_A$ is satisfied, the sentence matches the guiltiness if and only if the number of rational and honest agents is larger than the number of adversarial agents. Such an event happens with the following probability.

$$p^* \coloneqq \sum_{k=0}^{\lfloor (n-1)/2 \rfloor} \binom{n-1}{k} \delta_A^k (1 - \delta_A)^{(n-1)-k}.$$

If we have either (i) $\delta_A \approx 0$ or (ii) large n, then p^* is close to 1, while it is not exactly equal to 1 as long as $\delta_A > 0$.

Conversely, if we are more likely to have adversarial agents than honest agents (i.e., $\delta_A > \delta_H$), then Design 2 fails to punish guilty agents — in the unique equilibrium, the digital court acquits guilty agents with a large probability. Intuitively, this is because rational agents

cannot believe in honesty of the other agents, and therefore, the expected incentive payment is larger if they tell a lie. Accordingly, Design 2 can be used for enforcement of the agreement only if participants widely believe that honest behaviors are more likely to occur in Step 4+.

6.6 Example: Quadratic Psychological Cost

With general psychological cost functions, we can only characterize the equilibrium outcome in the limit of $\eta \to 0$. On the other hand, if we assume a specific functional form, we can characterize the equilibrium messages for general η. In this subsection, we assume that all psychological cost functions are quadratic and exhibit the equilibrium messages of rational, honest, and adversarial agents.

Theorem 4. *Suppose Assumption 1, either $\delta_H > 0$, and $\delta_A > 0$, and the following functional forms:*

$$c_H^1(m_j^i) = \lambda_H \cdot \left(1 - m_j^i\right)^2, \tag{7}$$
$$c_H^0(m_j^i) = \lambda_H \cdot \left(m_j^i\right)^2, \tag{8}$$
$$c_A^1(m_j^i) = \lambda_A \cdot \left(m_j^i\right)^2, \tag{9}$$
$$c_A^0(m_j^i) = \lambda_A \cdot \left(1 - m_j^i\right)^2, \tag{10}$$

where $\lambda_H, \lambda_A > 0$ represents the preference intensity of honest and adversarial agents, respectively. Then, the following reporting strategy profile σ constitutes the unique Bayesian Nash equilibrium: For every $i, j \in I$ and every $\omega \in \Omega$,

$$\sigma_j^i(\omega, R) = \begin{cases} \dfrac{\delta_H \lambda_H (\eta + \lambda_A)}{\delta_H \lambda_H (\eta + \lambda_A) + \delta_A \lambda_A (\eta + \lambda_H)} & \text{if } \omega_i = 1, \\ \dfrac{\delta_A \lambda_A (\eta + \lambda_H)}{\delta_H \lambda_H (\eta + \lambda_A) + \delta_A \lambda_A (\eta + \lambda_H)} & \text{if } \omega_i = 0, \end{cases}$$

$$\sigma_j^i(\omega, H) = \begin{cases} \dfrac{\eta}{\eta + \lambda_H} \cdot \dfrac{\delta_H \lambda_H (\eta + \lambda_A)}{\delta_H \lambda_H (\eta + \lambda_A) + \delta_A \lambda_A (\eta + \lambda_H)} + \dfrac{\lambda_H}{\eta + \lambda_H} & \text{if } \omega_i = 1, \\ \dfrac{\eta}{\eta + \lambda_H} \cdot \dfrac{\delta_A \lambda_A (\eta + \lambda_H)}{\delta_H \lambda_H (\eta + \lambda_A) + \delta_A \lambda_A (\eta + \lambda_H)} & \text{if } \omega_i = 0, \end{cases}$$

$$\sigma_j^i(\omega, A) = \begin{cases} \dfrac{\eta}{\eta+\lambda_A} \cdot \dfrac{\delta_H\lambda_H(\eta+\lambda_A)}{\delta_H\lambda_H(\eta+\lambda_A)+\delta_A\lambda_A(\eta+\lambda_H)} & \text{if } \omega_i = 1, \\ \dfrac{\eta}{\eta+\lambda_A} \cdot \dfrac{\delta_A\lambda_A(\eta+\lambda_H)}{\delta_H\lambda_H(\eta+\lambda_A)+\delta_A\lambda_A(\eta+\lambda_H)} + \dfrac{\lambda_A}{\eta+\lambda_A} & \text{if } \omega_i = 0. \end{cases}$$

Proof. Again, we focus on trial i and the case in which defendant i is innocent. We substitute (7), (8), (9), and (10) into (4), (5), and (6) and solve the equation system to obtain the unique equilibrium message profile, which are shown in Theorem 4. □

Rational agents vote for a correct decision if and only if

$$\delta_H\lambda_H(\eta+\lambda_A) > \delta_A\lambda_A(\eta+\lambda_H) \tag{11}$$

If there exists η with which (11) is satisfied, we can induce correct judgment with probability p^*.

Since (11) is linear in η, to check the existence of η that satisfies (11), it suffices to consider two cases: (i) $\eta \to 0$ and (ii) $\eta \to \infty$. The former case is considered in Theorem 2, and in such a case, (11) becomes equivalent to $\delta_H > \delta_A$.

As $\eta \to \infty$, (11) converges to $\delta_H\lambda_H > \delta_A\lambda_A$. Here, whether Design 2 can induce voting for a correct decision depends not only the relative population of honest agents δ_H, δ_A, but also the preference intensity, λ_H, λ_A. This is because the preference intensity influences the extent to which behavioral agents stick to their bliss points. Even when there is only a small fraction of honest agents, if they stick to truthful reporting, adversarial agents will compromise and report a message closer to the truth. Accordingly, the equilibrium messages become closer to the truthful one.

Note that, in the limit of $\eta \to \infty$, the equilibrium messages reported by each behavioral type converge to the same one. We have

$$\sigma_j^i(\omega, R), \sigma_j^i(\omega, H), \sigma_j^i(\omega, A) \to \begin{cases} \dfrac{\delta_H\lambda_H}{\delta_H\lambda_H+\delta_A\lambda_A} & \text{if } \omega_i = 1, \\ \dfrac{\delta_A\lambda_A}{\delta_H\lambda_H+\delta_A\lambda_A} & \text{if } \omega_i = 0, \end{cases}$$

as $\eta \to \infty$. This is because as η increases, intrinsic preferences become less important for minimization of behavioral agents' intrinsic preferences. Accordingly, when $\delta_H\lambda_H > \delta_A\lambda_H$ is satisfied, by increasing the scale of incentive payments (η), we can make adversarial agents to vote for a correct judgment. In this case, the decision is made unanimously with probability one, as opposed to the case of $\eta \to 0$.

However, we may need an extremely large η to achieve unanimity. To simplify calculation, let us focus on the case of $\lambda \coloneqq \lambda_H = \lambda_A$. In such a case, $\delta_H \lambda_H > \delta_A \lambda_H$ reduces to $\delta_H > \delta_A$, and adversarial agents vote for a correct decision if and only if

$$\frac{\eta}{\eta + \lambda} \cdot \frac{\delta_H}{\delta_H + \delta_A} > \frac{1}{2},$$

or equivalently,

$$\eta > \frac{\delta_H + \delta_A}{\delta_H - \delta_A} \cdot \lambda.$$

Accordingly, the scale of incentive payments, η, should be always larger than the scale of psychological cost, λ. Further more, when $\delta_H \approx \delta_A$, η becomes extremely large.

A large scale of payments is not desirable because (i) agents' worst-case loss (which happens off the equilibrium path) becomes large, and (ii) agents must make a large deposit into the digital court in Step 1+; thus the limited liability problem becomes severer. In this sense, Design 2 is not suitable to lead agents to unanimity, even if it is possible under some parameter values.

In Section 8, we consider an alternative design that also incentivizes agents to decide the sentence unanimously. Unlike in the case of Design 2 with large η, the scale of payments need not be larger than the scale of psychological cost by much.

7 Applications and Social Impacts

7.1 Auctions

This subsection demonstrates how untrusted sellers and buyers may run an auction as a self-enforcing decentralized mechanism using Design 2 as a digital court enforcing the contract.

For simplicity, we consider a second-price auction for a single indivisible good. The good is not digital (its ownership is not managed by a blockchain), and therefore, smart contracts cannot directly handle its ownership. There are three agents: agent 0 is a seller, and agent 1 and 2 are buyers. The seller's valuation for the auctioned good is assumed to be public and fixed to θ_0. This θ_0 is also a reservation price of the auction. Each buyer i has a private type $\theta_i \in [0, \bar{\theta}]$, which represents buyer i's valuation for the auctioned good. Each buyer i expresses her valuation θ_i to the seller by sending a bid $\xi_i \in [0, \bar{\theta}]$, and the buyer who made a higher bid will win. For simplicity, we ignore the case of ties. Then, the buyer will pays the loser's bid or the reservation price to the seller.

Initially, these three agents interact and make an agreement α. In this application, α specifies the rule of the auction. The seller 0's action space is $A_0 \coloneqq \{0, 1, 2\}$, where a_0

indicates the recipient of the good ($a_0 = 0$ means the seller keeps it). Each buyer $i \in \{0, 1\}$'s action space is $A_i = [0, \bar{\theta}]$, where a_i represents the amount of money transferred from buyer i to the seller. Specifically, α is given by

$$\alpha_0(\xi) = \begin{cases} 0 & \text{if } \max\{\theta_0, \xi_1, \xi_2\} = \theta_0, \\ 1 & \text{if } \max\{\theta_0, \xi_1, \xi_2\} = \xi_1, \\ 2 & \text{if } \max\{\theta_0, \xi_1, \xi_2\} = \xi_2. \end{cases}$$

$$\alpha_1(\xi) = \begin{cases} \max\{\theta_0, \xi_2\} & \text{if } \max\{\theta_0, \xi_1, \xi_2\} = \xi_1, \\ 0 & \text{otherwise.} \end{cases}$$

$$\alpha_2(\xi) = \begin{cases} \max\{\theta_0, \xi_1\} & \text{if } \max\{\theta_0, \xi_1, \xi_2\} = \xi_2, \\ 0 & \text{otherwise.} \end{cases}$$

Then, an agent (typically the seller) deploys Design 2 of the digital court on a blockchain. The seller deposits $\bar{t}_0 = \theta_0 + 3\eta$, and each buyer deposits $\bar{t}_i = \bar{\theta} + 3\eta$ to it. After that, each buyer i simultaneously announces her bid ξ_i. At this point, all agents become aware of the outcome of the auction. Note that bidding need not be processed on a blockchain. In particular, if agents do not want to disclose the information about bidding to the public, they should process the bidding step privately, without using a blockchain-based smart contract.

Once agents agree on the auction outcome, they execute it. The seller delivers the object in accordance with $\alpha_0(\xi)$, and each buyer makes a payment in accordance with $\alpha_i(\xi)$. Of course, each agent can renege on the agreement — the seller can keep holding the auctioned good (choose $\hat{a}_0 = 0$), and the buyers can refuse to make a payment (choose $\hat{a}_i = 0$). Hence, agents cannot run an auction without using some enforcement.

After every agent takes an action $\hat{a}_i$, she verifies her action to all the other agents. By comparing $\hat{a}$ and $\alpha(\xi)$, all agents figure out who violated the agreement. Then, agents simultaneously submit a message m_i to the digital court. The digital court extracts $\gamma_i(m)$ from agent i's deposit and repays $\bar{t}_i - \gamma_i(m)$ to agent i.

Seller 0's resultant payoff is

$$\hat{a}_1 + \hat{a}_2 - \theta_0 \cdot \mathbb{1}\{\hat{a}_0 = 0\} - \gamma_0(m).$$

Buyer $i \in \{1, 2\}$'s resultant payoff is

$$\theta_i \cdot \mathbb{1}\{\hat{a}_0 = i\} - \hat{a}_i - \gamma_i(m).$$

In the unique Bayesian Nash equilibrium, γ_i specified by Definition 5 identifies the set of guilty agents and fines (slightly more than) the maximum possible again from deviation. Accordingly, in a perfect Bayesian equilibrium, all agents $i \in \{0, 1, 2\}$ takes the agreed action, $\alpha_i(\xi)$.

7.2 Bidding Ring

Parties can also use the digital court for an illegal purpose. For example, parties can use the digital court to form a *strong cartel*, in which the cartel members can make transfer payments (McAfee and McMillan 1992). Transfers within a cartel are often prohibited by competition by law; thus, legal enforcement is not available for profit reallocation. Therefore, thus far, repeated interaction among cartel members has been necessary to incentivize (i) the winner to share the profit with the losers, and (ii) the losers to lower the bid so as to reduce the price paid by the winner. However, using a digital court as a commitment device, parties can form a strong cartel even when they interact only once.

In this subsection, we demonstrate that bidders can extract full surplus from a first-price auction with common values. Note that first-price auctions are considered to be more robust against collusion than the other auctions, such as second-price auctions, etc.[12] There are two bidders, $I := \{1, 2\}$, whose value of the auctioned good is $\bar{\theta} \in (0, 1]$. The value is common knowledge among agents. The seller is not modeled as a strategic agents, but holds a first-price auction. The reservation price of the auction is zero, and a tie is broken equally at random. In a competitive environment, in the unique Bayesian Nash equilibrium, both bidders bid $\bar{\theta}$ and the good is sold at the price of $\bar{\theta}$; thus, the *seller* extracts full surplus.

Each bidder i's action a_i is a pair of (i) her bidding decision in the first-price auction, $d_i \in [0, 1]$ and (ii) the monetary payment to the other party, $p_i \in [0, 1]$. Since agents have no private information, they need not exchange messages to decide the action profile. The agreement specifies that each bidder i (i) submits $d_i = 0$ to the first-price auction, and (ii) transfers $p_i = \bar{\theta}/2$ to the loser if she wins the good (this happens with probability a half if both agents bid zero).

Whenever a bidder wins the object, she is supposed to share the profit with the loser. Thus, a bidder cannot increase her profit just by increasing her bid, d_i. The only profitable deviation is to refuse to make a payment to the loser. Here, the maximum gain from a deviation is $\bar{\theta}/2$; thus, if each bidder initially deposits $\bar{\theta}/2 + \epsilon$ to the digital court, then the digital court can offset incentives for deviation. Here, the price of the good is zero; thus, bidders steal all the surplus.

[12] See Chapter 11 of Krishna (2009), for example.

Such a scheme is applicable for not only first-price auctions but also many other classes of auctions. Furthermore, bidders can extract full surplus by developing a similar agreement even in the case of private values (see Section 6 of McAfee and McMillan 1992). Although the seller can mitigate this problem by introducing a random reservation price and anonymizing the identity of buyers, typically bidders can still increase their joint profits by forming a strong cartel.

7.3 Regulation

Smart contracts enable parties to use self-enforcing mechanisms, which are aimed for either a legitimate purpose (e.g., auctions) or illegal purpose (e.g., bidding rings). This feature is contrasting to traditional mechanism design with legal enforcement — since courts always check the legality of the contract to be enforced, parties cannot leverage legal enforcement for running mechanisms intended for illegal objectives. Ideally, a regulator wants to permit a legitimate use of smart contracts but prohibit an illegal use. However, practical implementation of such regulation seems a challenging problem.

Public blockchains disclose all the information recorded there to any party. Accordingly, smart contracts deployed on a public blockchain is also disclosed to the public. By observing the digital court, a regulator can figure out that some parties are attempting to enforce a certain agreement.

If the regulator can also find that the parties who involved in the contract are not allowed to make a privacy-preserving binding agreement (e.g., parties are construction companies competing in procurement), then it is enough to expose the fact that the parties uploaded a digital court on a blockchain. To implement this regulation, the regulator must be able to identify each party's pseudonym (account number) in the blockchain system. Although the identity of users can be inferred from the history of transactions, whether real-world regulators can always detect the pseudonym of each party or not is ambiguous.

The problem is severer if parties are allowed to make privacy-preserving agreements that are intended for legitimate purposes. In such a case, detecting the pseudonym is not sufficient to expose a crime — the regulator must verify that the agreement enforced by the digital court is illegal. However, since the information about the agreement (α) need not be written on the smart contract, the regulator cannot detect whether the agreement is illegal or not just by checking the information that appears on the blockchain.

Hence, if contracting itself is legal, then the regulator must investigate the contract details by searching for evidence in actions taken in the real world, rather than analyzing the transaction data recorded in the blockchain. Seeking evidence for a contract with blockchain

enforcement is more difficult than the case of illegal activities based on repeated interactions. In the case of repeated interactions, the regulator can investigate the time-series data of agents' actions (e.g., the history of bidding in a series of auctions). Here, the regulator can apply a statistical analysis to expose the pattern of illegal activities. In contrast, blockchain enforcement does not need a long-term relationship, and therefore, may be used for enforcing a one-shot agreement. If the interaction is one-shot, the regulator may not be able to detect illegal activities statistically.

Furthermore, some illegal agreements leave no decisive evidence in the real world. To incentivize agents to take collusive actions, it often suffices to promise that winners of the competition distribute some profits to losers. Such transfers need not be taken explicitly — if the winner and loser have a business relationship, the winner can transfer a profit by giving a discount in other deal, for example. Although the outside parties cannot distinguish such transfers from a legitimate dealing, the winner and loser can secretly agree to interpret the discount as a reward for taking collusive actions. If the incentive for taking collusive actions and making a transfer from winners to losers is provided by a digital court, there is no decisive evidence for a regulator to expose the collusive agreement.

There is some hope that the digital-court approach is inconvenient for criminal activities. A digital court takes advantage of the chance to have honest agents, who truthfully report the state to the smart contract. If the psychological motivation for "honesty" comes from moral consciousness, it might be awoken only when the mechanism is used for a good purpose. If criminal parties have a sense of guilt and have a fear and doubt about the other agents' messaging (i.e., suspect that the other agents might be adversarial with a relatively high probability), then unique implementation might not be achievable. The regulator might also be able to take advantage of this nature so as to disturb illegal self-enforcing mechanisms.

8 False Charge Problem

8.1 Preventing False Charges in Sentences

When the realized number of adversarial jurors is large, Design 2 may "convict" innocent defendants because adversarial jurors vote for an incorrect decision. Although the probability that an innocent defendant is convicted is small (as long as the ex ante fraction of adversarial agents is small and the total number of agents is large), such an event occurs with a positive probability. This is a *false charge*. This section considers an alternative design of a mechanism that never convicts innocent defendants in the unique equilibrium. We construct a mechanism that encourages adversarial agents to tell the truth.

Design 3 is a hybrid of Design 1 and 2 — Design 3 asks agents to submit *both* a continuous message $m_j^i(1) \in [0,1]$ and a binary message $m_j^i(2) \in \{0,1\}$, simultaneously. As each agent submits multiple messages to each trial, Assumption 1 is not directly applicable. Here, we introduce a slightly different assumption of intrinsic preferences.

Assumption 2. For every agent $i, j \in I$, $M_j^i = M_j^i(1) \times M_j^i(2)$, where $M_j^i(1) \subseteq [0,1]$ and $M_j^i(2) \subseteq \{0,1\}$. For every agent $j \in I$, with probability $1 - \delta_H - \delta_A$, agent j is *rational* ($b_j = R$) and wants to minimize

$$\Gamma_j(m, \omega, R) \coloneqq \gamma_j(m).$$

With probability $\delta_H > 0$, agent j is *honest* ($b_j = H$) and wants to minimize the following function:

$$\begin{aligned} &\Gamma_j(m, \omega, H) \\ \coloneqq &\gamma_j(m) + \sum_{i \in I} \left\{ \omega_i \cdot c_H^1(m_j^i(1)) + (1 - \omega_i) \cdot c_H^0(m_j^i(1)) + \lambda_H \cdot \mathbb{1}\left\{ m_j^i(2) \neq \omega_i \right\} \right\}, \end{aligned}$$

where (i) both $c_H^1 : [0,1] \to \mathbb{R}_+$ and $c_H^0 : [0,1] \to \mathbb{R}_+$ are strictly convex and twice differentiable, (ii) c_H^1 is strictly decreasing and $c_H^1(1) = 0$, (iii) c_H^0 is strictly increasing and $c_H^0(0) = 0$, and (iv) $\lambda_H > 0$. With probability $\delta_A \geq 0$, agent j is *adversarial* ($b_j = A$) and wants to minimize the following function:

$$\begin{aligned} &\Gamma_j(m, \omega, A) \\ \coloneqq &\gamma_j(m) + \sum_{i \in I} \left\{ \omega_i \cdot c_A^1(m_j^i(1)) + (1 - \omega_i) \cdot c_A^0(m_j^i(1)) + \lambda_A \cdot \mathbb{1}\left\{ m_j^i(2) = \omega_i \right\} \right\}, \end{aligned}$$

where (i) both $c_A^1 : [0,1] \to \mathbb{R}_+$ and $c_A^0 : [0,1] \to \mathbb{R}_+$ are strictly convex and twice differentiable, (ii) c_A^1 is strictly increasing and $c_A^1(1) = 0$, (iii) c_A^0 is strictly decreasing and $c_A^0(0) = 0$, and (iv) $\lambda_A > 0$.

The realization for each agent is independent, and each agent cannot observe whether other agent is honest or not.

Parallel to Assumption 1, honest agents incur a larger psychological cost as they submit less truthful messages, and adversarial agents incur a larger psychological cost as they submit more truthful messages. The psychological cost for the first message, $m_j^i(1)$, is specified in the same manner as Assumption 1. Since the second message space is binary, without loss of generality, we can assume that each behavioral agent incurs a psychological cost of λ_H or λ_A when she reports a dispreferred message.

Design 3 is defined as follows.

Definition 6. *Design 3* of the digital court is specified as follows:

- $M_j^i \coloneqq M_j^i(1) \times M_j^i(2)$, where $M_j^i(1) \coloneqq [0,1]$ and $M_j^i(2) \coloneqq \{0,1\}$ for all $i, j \in I$.
- Defendant i's sentence function s^i is given by

$$s^i(m_{-i}^i) \coloneqq \begin{cases} 1 & \text{if } \sum_{j \neq i} m_j^i(2) > \dfrac{n-1}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

- Juror j's incentive payment term for trial i, q_j^i, is given by

$$q_j^i(m^i) \coloneqq \frac{\eta(1)}{n-1} \cdot \sum_{k \neq j} \left(m_j^i - m_k^i\right)^2 + \eta(2) \cdot \mathbb{1}\left\{m_j^i(2) \neq \mu_{-j}^i\left(m_{-j}^i(1)\right)\right\},$$

  where $\eta(1), \eta(2) > 0$ and

$$\mu_{-j}^i(m_{-j}^i(1)) \coloneqq \begin{cases} 1 & \text{if } \sum_{k \neq j} \mathbb{1}\left\{m_k^i(1) > \dfrac{1}{2}\right\} > \dfrac{n-1}{2}, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

- Agent j's deposit $\bar{t}_j$ is given by

$$\bar{t}_j \coloneqq T_j + \eta(1) + \eta(2).$$

As in Design 2, the message space of the first message $m_j^i(1)$ is $[0,1]$. The structure of the payment rule for jurors is also the same: Design 3 compares $m_j^i(1)$ and $m_k^i(1)$ for every j, k, and fines $[\eta(1)/(n-1)] \cdot (m_j^i(1) - m_k^i(1))^2$. Except this term, $m_j^i(1)$ does not affect agent j's payment. Hence, agent j's incentive for reporting $m_j^i(1)$ in Design 3 is the same as in Design 2. Accordingly, as shown in Theorem 2, if we have $\delta_H > \delta_A$ and $\eta \to 0$, then all rational and honest agents vote for a correct decision. From now, we focus on such a case.

If η is small, adversarial agents ignore material payoffs and always tell a lie. Hence, the first message profile is typically not unanimous. If a digital court decides the sentence using $m^i(1)$, with a positive probability, it misjudges (as Design 2 does). To avoid this, Design 3 require jurors to make one more step of peer prediction, and the second message is used for deciding the sentence. Specifically, Design 3 compares juror j's second message, $m_j^i(2)$, with

the majority opinion of the first messages of all the other agents, $\mu^i_{-j}(m^i_{-j}(1))$. If these two do not coincide, Design 3 fines juror j by $\eta(2)$.

Note that, the functional form of the majority opinion function of Design 3 is the same as the sentence function of Design 2. Hence, if $\delta_H > \delta_A$ and η is small, then the probability that the majority opinion $\mu^i_j(m^i_{-j}(1))$ is correct is also p^*.

Under Design 3 with appropriate choice of $\eta(1)$ and $\eta(2)$, in the unique Bayesian Nash equilibrium, the decision is made unanimously for every trial. Hence, the digital court never misidentifies the set of guilty agents.

Theorem 5. *Suppose Assumption 2 and $\delta_H > \delta_A$. Then, for any $c^1_H, c^0_H, c^1_A, c^0_A, \lambda_H, \lambda_A$, there exist $\eta(1)$ and $\eta(2)$ with which Design 3 satisfies the following conditions:*

1. *Design 3 becomes dominance solvable and has a unique Bayesian Nash equilibrium.*
2. *In the unique Bayesian Nash equilibrium, the judgment is correctly made; i.e., $s^i(m^i_{-i}) = \omega_i$ for all $i \in I$. Furthermore, the decision is made unanimously in the sense that every rational, honest, and adversarial agent $j \in I$ reports $m^i_j(2) = \omega_i$ for all $i \in I$.*

Proof. Again, we focus on trial i and assume that defendant i is innocent; i.e., $\omega_i = 0$. Parallel to Theorem 2, in the limit of $\eta(1) \to 0$, all honest agents report $m^i_j(1) = 0$, all adversarial agents report $m^i_j(1) = 0$, and all rational agents report $m^i_j = \delta_H/(\delta_H + \delta_A) < 1/2$. Hence, if we take a sufficiently small $\eta(1)$, then (i) rational and honest agents report $m^i_j < 1/2$, and (ii) adversarial agents report $m^i_j > 1/2$. We pick such $\eta(1)$. Then, the probability that the majority opinion μ^i_{-j} is correct is p^*.

Now, we consider adversarial agent j's incentive for reporting the second message. If an adversarial juror j tells a lie, her expected fine is

$$p^* \cdot \eta(2). \tag{13}$$

On the other hand, the total of her expected fine and psychological cost is

$$(1 - p^*) \cdot \eta(2) + \lambda_A. \tag{14}$$

If we take $\eta(2) > \lambda_A/(2p^* - 1)$, then (13) becomes larger than (14). Accordingly, all adversarial agents send a truthful second message to the digital court in the unique equilibrium. We can also show that all the rational and honest agents tell the truth in a similar manner. □

The first message is for constructing a reference point. Just like Design 2, Design 3 allows agents to make fractional voting, and therefore, (under a certain assumption) agents vote

for a correct decision with a large probability. Design 3 does not use it as the final decision but require agents to make one more report. The second message is used for deciding the sentence. Just like Design 1, this message is binary; thus, if an agent tells a lie, then the message becomes very distant from the majority opinion. This feature increases the effectiveness of the punishment.

Note that Design 3 needs a relatively larger scale of the incentive payments compared with the case of Design 2 with $\eta \to 0$. Since $\eta(2)$ should be larger than $\lambda_A/(2p^*-1)$, Design 3 requires at least $T_j+\lambda_A/(2p^*-1)$ as a deposit, and each juror indeed loses $\eta(2) > \lambda_A/(2p^*-1)$ when $m_j^i(2)$ does not coincide with $\mu_{-j}^i(m_{-j}^i(1))$. This is because we need to incentivize adversarial agents to tell the truth in order to make correct judgment with probability one — the incentive payment terms must offset adversarial agents' psychological payoff from telling a lie. However, the required payment scale is typically much smaller than the case of Design 2 with large η (demonstrated in Subsection 6.6).

8.2 Preventing False Charges in Jurors' Incentive Payments

Design 3 still has a weak point. In Design 3, although all the adversarial agents make truthful reporting for the second message, they do not tell the truth for the first message. Hence, the majority opinion of the first messages is incorrect with probability $1-p^*$. Accordingly, even if an agent correctly reports $m_j^i = \omega_i$, she will be fined $\eta(2) > \lambda_A/(2p^*-1)$ with probability $1-p^*$, due to inconsistency between her second message and the majority opinion of the first messages. This subsection extends Design 3 to decrease the incentive payment imposed on agents who vote for a correct decision.

As our assumption on intrinsic preferences crucially depend on the specification of the mechanism, we first define the mechanism, Design 4. Design 4 is similar to Design 3, but it requires many ($Z \geq 3$) messages simultaneously.

Definition 7. *Design 4* of the digital court is specified as follows:

- $M_j^i \coloneqq \prod_{z=1}^{Z} M_j^i(z)$, where $M_j^i(1) \coloneqq [0,1]$ and $M_j^i(z) \coloneqq \{0,1\}$ for all $z \geq 2$, for all $i,j \in I$.
- To decide the sentence for defendant i, Design 4 initially draws z from $\{2,3,\dots,Z\}$ equally at random. Then, the sentence for defendant i is determined by a simple

majority rule with respect to $m^i_{-i}(z)$:

$$S^i_z(m^i_{-i}(z)) \coloneqq \begin{cases} 1 & \text{if } \sum_{j\neq i} m^i_j(z) > \dfrac{n-1}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

The probability that S^i_z is chosen as the resultant sentence function is

$$\text{Prob}\left(s^i(m^i_{-i}) = S^i_z(m^i_{-i}(z))\right) = \frac{1}{Z-1}$$

for all $z = 2, 3, \ldots, Z$.

- Juror j's incentive payment term for trial i, q^i_j, is given by

$$q^i_j(m^i) \coloneqq \frac{\eta(1)}{n-1} \cdot \sum_{k\neq j} (m^i_j(1) - m^i_k(1))^2 + \eta(2) \cdot \mathbb{1}\left\{m^i_j(2) \neq \mu^i_{-j}(m^i_{-j}(1))\right\} + \sum_{z=3}^{Z} \eta(z) \cdot \mathbb{1}\left\{m^i_j(z) \neq \zeta^i_{-j}(m^i_{-j}(z-1))\right\},$$

where μ^i_{-j} is given by (12) and ζ^i_{-j} is given by

$$\zeta^i_{-j}(m^i_{-j}(z)) \coloneqq \begin{cases} 1 & \text{if } \sum_{k\neq j} m^i_k(z) > \dfrac{n-1}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

- Agent j's deposit $\bar{t}_j$ is given by

$$\bar{t}_j \coloneqq T_j + \sum_{z=1}^{Z} \eta(z).$$

Design 4 is a random mechanism. Each agent submits $Z \geq 3$ messages, and Design 4 chooses z from 2 to Z uniformly at random. The message profile corresponds to the realized z, $m^i_{-i}(z)$, is used for deciding the sentence for agent i. Note that, Design 4 involves randomness only for this part, and agents' incentive payments, q^i_j, are deterministic. Hence, as for derivation of the Bayesian Nash equilibrium, we can ignore the randomness. (We formulate Design 4 as a random mechanism for justifying the specification of the psychological cost in Assumption 3.)

In Design 4, the first message, $m_j^i(1)$ is very special in the sense that (i) its space is an interval $[0,1]$, while all the other message spaces are binary, $\{0,1\}$, and (ii) the first message is not used for deciding the sentence. All the other messages have a similar role — they are binary, and all the agents are expected to match their messages $m_j^i(z)$ with the guiltiness of the defendant, ω_i. Moreover, all these messages are used for deciding the sentence with an equal probability.

Reflecting this design intention of Design 4, we specify intrinsic preferences of honest and adversarial agents in the following manner.

Assumption 3. For every $i, j \in I$, $M_j^i \coloneqq \prod_{z=1}^{Z} M_j^i(z)$, where $Z \geq 3$. $M_j^i(1) \subseteq [0,1]$ and $M_j^i(z) \subseteq \{0,1\}$ for all $z \geq 2$. For every agent $j \in I$, with probability $1 - \delta_H - \delta_A$, agent j is *rational* ($b_j = R$) and wants to minimize

$$\Gamma_j(m, \omega, R) \coloneqq \gamma_j(m).$$

With probability $\delta_H \geq 0$, agent j is *honest* ($b_j = H$) and wants to minimize the following function:

$$\begin{aligned} &\Gamma_j(m, \omega, H) \\ &\coloneqq \gamma_j(m) + \sum_{i \in I} \left\{ \omega_i \cdot c_H^1(m_j^i(1)) + (1 - \omega_i) \cdot c_H^0(m_j^i(1)) + \frac{\lambda_H}{Z-1} \cdot \sum_{z=2}^{Z} \mathbb{1}\left\{ m_j^i(z) \neq \omega_i \right\} \right\}, \end{aligned}$$

where (i) both $c_H^1 : [0,1] \to \mathbb{R}_+$ and $c_H^0 : [0,1] \to \mathbb{R}_+$ are strictly convex and twice differentiable, (ii) c_H^1 is strictly decreasing and $c_H^1(1) = 0$, (iii) c_H^0 is strictly increasing and $c_H^0(0) = 0$, and (iv) $\lambda_H > 0$.

With probability $\delta_A \geq 0$, agent j is *adversarial* ($b_j = A$) and wants to minimize the following function:

$$\begin{aligned} &\Gamma_j(m, \omega, A) \\ &\coloneqq \gamma_j(m) + \sum_{i \in I} \left\{ \omega_i \cdot c_A^1(m_j^i(1)) + (1 - \omega_i) \cdot c_A^0(m_j^i(1)) + \frac{\lambda_A}{Z-1} \cdot \sum_{z=2}^{Z} \mathbb{1}\left\{ m_j^i(z) = \omega_i \right\} \right\}, \end{aligned}$$

where (i) both $c_A^1 : [0,1] \to \mathbb{R}_+$ and $c_A^0 : [0,1] \to \mathbb{R}_+$ are strictly convex and twice differentiable, (ii) c_A^1 is strictly increasing and $c_A^1(0) = 0$, (iii) c_A^0 is strictly decreasing and $c_A^0(1) = 0$, and (iv) $\lambda_A > 0$.

The realization for each agent is independent, and each agent cannot observe whether other agent is honest or not.

Similar to Assumption 2, Assumption 3 introduces $c_H^1, c_H^0, c_A^1, c_A^0$ as a psychological cost function for the first message. Assumption 3 is a generalization of Assumption 2 in the sense that if we take $Z = 2$, then Assumption 3 becomes identical to Assumption 2.

In Assumption 3, for every time a behavioral agent sends a dispreferred message, she incurs a (expected) psychological cost of $\lambda_H/(Z-1)$ or $\lambda_A/(Z-1)$. The following is an interpretation of this specification. These $Z-1$ message profiles are selected for deciding the sentence with an equal probability, $1/(Z-1)$. Importantly, when a message is not selected, it has no influence on the sentence to be made. As a message becomes a "white lie" in such a case, it is natural to assume that each agent may incur a psychological cost only if the relevant message is used for deciding the sentence. We assume that an agent incurs an *ex post* psychological cost of λ_H or λ_A when her dispreferred message happens to be used to decide the sentence. Then her *expected* psychological cost of sending a dispreferred message for one time should be equal to $\lambda_H/(Z-1)$ or $\lambda_A/(Z-1)$.

Now, let us consider juror j's incentive of reporting. Using the first message $m_j^i(1) \in [0,1]$, Design 4 constructs a reference point, $\mu_{-j}^i(m_{-j}^i(1))$. Then, each juror j predicts the majority opinion of the other jurors by sending a binary message, $m_j^i(2) \in \{0,1\}$. If $m_j^i(2)$ and $\mu_{-j}^i(m_{-j}^i(1))$ do not match, juror j is fined $\eta(2)$. Thus far, the construction of incentive payments in Design 4 is identical to that in Design 3.

Design 4 has an additional payment rule. By reporting z-th message, $m_j^i(z)$, each juror j repeatedly predicts the majority opinion of the other agents with respect to the preceding message profile, $\xi_{-j}^i(m_{-j}^i(z-1))$. Each juror does this for $Z-2$ times by reporting $m_j^i(z)$ for $z = 3, 4, \ldots, Z$. The same as the second message, each agent's message space for the third message or later is binary. If a juror fails to predict it with z-th report, she will be fined $\eta(z)$.

The majority opinion of later messages (the second messages of after) is a more accurate predictor than the majority opinion of the first message. The majority opinion of the first message is incorrect with probability $1-p^*$. Hence, if we impose agents a fine based on the comparison against the first message profile, we would cause a false charge. On the other hand, since $m_j^i(2), \ldots, m_j^i(Z)$ are chosen from a binary message space, and all agents submit an exactly correct message with probability one, if we punish agents based on comparison against later message profiles, then we no longer have a problem of the false charge. The only "false charge" imposed by Design 4 is from the comparison between $m_j^i(2)$ and $\mu_{-j}^i(m_{-j}^i(1))$. As we increase the number of messages, Z, we can make the false charge incurred by this comparison smaller.

Just like Design 3, Design 4 correctly identifies the set of guilty agents with probability one. Furthermore, the false charge paid by honest and rational jurors vanishes as $Z \to \infty$.

Theorem 6. *Suppose Assumption 2 and $\delta_H > \delta_A$. Then, for any c_H^1, c_H^0, c_A^1, c_A^0, λ_H, λ_A, and $\epsilon > 0$ there exists $(\eta(z))_{z=1}^Z$ with which Design 4 satisfies the following conditions:*

1. *Design 4 is dominance solvable and has a unique Bayesian Nash equilibrium.*
2. *In the unique Bayesian Nash equilibrium, the judgment is correctly made; i.e., $s^i(m^i_{-i}) = \omega_i$ for all $i \in I$ happens with probability one. Furthermore, the decision is made unanimously in the sense that every rational, honest, and adversarial agent $j \in I$ reports $m^i_j(z) = \omega_i$ for all $i \in I$ and $z = 2, \dots, Z$ with probability one.*
3. *In the unique Bayesian Nash equilibrium, the total fine imposed on an innocent agent is always smaller than*
$$n \cdot \frac{\lambda_A}{(2p^* - 1)(Z-1)} + \epsilon. \tag{15}$$
4. *Agent j's deposit $\bar{t}_j$ is smaller than*
$$T_j + n \cdot \lambda_A \cdot \left(1 + \frac{2(p^*+1)}{(2p^*-1)(Z-1)}\right) + \epsilon. \tag{16}$$

Proof. We focus on trial i. We first assume that defendant i is innocent; i.e., $\omega_i = 0$. Parallel to Theorem 5, if we take a small $\eta(1)$, then $\mu^i_{-j}(m^i_{-j}(1)) = \omega_i$ happens with probability p^*.

We first consider agent j's incentive for sending her second message, $m^i_j(2)$. If she tells a lie, the sum of her expected fine from the second message is

$$p^* \cdot \eta(2). \tag{17}$$

If she tells the truth, the total of her expected fine and psychological cost is

$$(1 - p^*) \cdot \eta(2) + \frac{\lambda_A}{Z-1}. \tag{18}$$

(18) is smaller than (17) if

$$\eta(2) > \frac{\lambda_A}{(2p^* - 1)(Z-1)}. \tag{19}$$

Therefore, if we take $\eta(2)$ that satisfies (19) for all agent j, then in any equilibrium, all agents tell the truth for the second message.

Suppose that all agents report $m^i_j(z'-1) = 0$ in any equilibrium. If so, $\zeta^i_{-j}(m^i_{-j}(z'-1)) = 0$ also holds with probability one in any equilibrium. We consider her incentive for reporting z'-th message. If she tells a lie, the sum of her expected fine from z'-th message is

$$\eta(z'). \tag{20}$$

On the other hand, if she tells the truth, then she is not fined, and her psychological cost is

$$\frac{\lambda_A}{Z-1}. \tag{21}$$

(21) is smaller than (20) if

$$\eta(z') > \frac{\lambda_A}{Z-1}.$$

Therefore, if we take $\eta(z')$ that satisfies (19), then in any equilibrium, all agents tell the truth for z'-th message.

By mathematical induction, we can prove that, if $\omega_i = 0$, in the unique equilibrium, all the adversarial jurors tell the truth: $m_j^i(2) = m_j^i(3) = \cdots = m_j^i(Z) = \omega_i$. We can prove the case of $\omega_i = 1$ in a similar manner.

Hence, taking $\eta(2), \ldots, \eta(Z)$ such that

$$\eta(2) > \frac{\lambda_A}{(2p^* - 1)(Z-1)}, \tag{22}$$

$$\eta(z') > \frac{\lambda_A}{Z-1}, \tag{23}$$

for all $z' = 3, \ldots, Z$, we can make all adversarial agents to tell the truth in any equilibria.

As rational and honest agents have stronger incentives to tell the truth, they also have an incentive to tell the truth. Accordingly, Design 4 has a unique Bayesian Nash equilibrium, and all agents report $m_j^i(z) = \omega_i$ for all $z = 2, \ldots, Z$ in the unique Bayesian Nash equilibrium.

In the unique Bayesian Nash equilibrium, an innocent agent may pay $\eta(1) + \eta(2)$ for trial j because all agents reporting after $z = 2$ always perfectly coincide. As $\eta(1)$ can be arbitrarily small, and the infimum of $\eta(2)$ is given by the right hand side of (22). Hence, we can take $(\eta(z))_{z=1}^Z$ with which $\eta(1) + \eta(2)$ is smaller than (15).

Furthermore, the total deposit required for Design 4 is $\bar{t}_j = T_j + \sum_{z=1}^Z \eta(z)$, and its infimum amount is obtained by summing up (22) and (23) for all $z' = 3, 4, \ldots, Z$. Hence, we can take $(\eta(z))_{z=1}^Z$ with which $T_j + \sum_{z=1}^Z \eta(z)$ is smaller than (16), as desired. □

In the limit of $Z \to \infty$, the infimum amount of the total fine imposed on an innocent agent, (15), converges to zero. Hence, Design 4 approximately resolves the false charge problem if we take a large Z. On the other hand, the infimum amount of each agent j's deposit, (16), converges to $T_j + n\lambda_A$. Accordingly, even in the limit of $Z \to \infty$, the deposit does not diverge.

9 Concluding Remarks

We study the design of self-enforcing mechanisms, using a smart contract as a commitment device. We design digital courts that replace the role of courts in the traditional mechanism design paradigm. Digital courts incentivize agents to report those who violated the agreement truthfully and impose fines to them. The punishment deters agents from reneging on the agreement, just as a court does in the traditional paradigm. Hence, any agreement that is implementable with legal enforcement can also be implemented with the presence of a platform for smart contracts.

To make correct judgment, the digital court has to incentivize agents to input truthful information (oracle problem). This paper studies unique implementation of the correct judgment as a mechanism design problem. As smart contracts can only execute monetary transfers, unique implementation is impossible if all agents are purely interested in material payoffs. Hence, we develop a mechanism that has a unique Bayesian Nash equilibrium if a positive (small) fraction of agents prefer to tell the truth or tell a lie (Design 2). Design 2 successfully incentivizes rational agents to vote for a correct decision if honest agents are more likely to exist than rational agents. Furthermore, the possibility of a false charge can be prevented by requiring agents to send multiple messages (Design 3, Design 4). Note that, the assumptions leveraged in Design 2, 3, and 4 can be generalized further. For example, as long as the psychological cost functions satisfy the assumptions on its shape (monotonicity and strict convexity), we can prove a similar possibility theorem even if the cost functions vary across agents.

Digital courts can be used for implementing any agreement. The detail of the agreement is not leaked even if a digital court is uploaded to a public blockchain. As our digital-court approach only utilizes the tools already practically implemented as of 2020, parties can form an enforcing agreement using our approach right now. If smart contracts and digital courts are used for a legitimate purpose, they would improve the social welfare by extending the possibility of mechanism design.

This result also implies that blockchain disruption might be more serious than the literature has expected. Before the creation of cryptocurrencies, it has been difficult for criminals to make a viable agreement because they could not rely on legal enforcement. The emergence of blockchains enabled parties (including criminals) to run self-enforcing mechanisms. Even if the regulator monitors the blockchain, the regulator cannot detect whether a digital court is used for a legitimate or unlawful purpose. Again, since the digital-court approach is already available, criminals might be utilizing a similar scheme to enforce illegal agreements even now. To protect society from this threat, the regulatory authority should carefully

consider a way to prevent parties from abusing smart contracts. The design of desirable regulations is an interesting future research topic.

Our digital-court approach crucially relies on the behavioral assumptions. Whether and how digital courts are used in the real world crucially depends on agents' preferences for honesty. Agents' intrinsic preferences must be influenced by various factors, such as the instruction of the digital court, the detail of the agreement (e.g., the legality of the agreement), and the attribute of party members. Whether agents find that the digital-court approach is useful and prefer it over the traditional legal enforcement would depend on contexts, and we can test the contexts in which the digital-court approach performs well through economic experiments and empirical analyses. However, these analyses are beyond the scope of this paper.

References

Abeler, J., D. Nosenzo, and C. Raymond (2019): "Preferences for truth-telling," *Econometrica*, 87, 1115–1153.

Abreu, D. and H. Matsushima (1992): "Virtual implementation in iteratively undominated strategies: complete information," *Econometrica*, 60, 993–1008.

——— (1994): "Exact implementation," *Journal of Economic Theory*, 64, 1–19.

Abreu, D. and A. Sen (1991): "Virtual implementation in Nash equilibrium," *Econometrica*, 997–1021.

Adler, J., R. Berryhill, A. Veneris, Z. Poulos, N. Veira, and A. Kastania (2018): "Astraea: A decentralized blockchain oracle," in *2018 IEEE Conferences on Internet of Things, Green Computing and Communications, Cyber, Physical and Social Computing, Smart Data, Blockchain, Computer and Information Technology, Congress on Cybermatics*, IEEE, 1145–1152.

Akbarpour, M. and S. Li (2019): "Credible Auctions: A Trilemma," *Econometrica*, forthcoming.

Antonopoulos, A. M. and G. Wood (2018): *Mastering Ethereum: building smart contracts and DApps*, O'Reilly Media.

Asgaonkar, A. and B. Krishnamachari (2019): "Solving the buyer and seller's dilemma: A dual-deposit escrow smart contract for provably cheat-proof delivery and

payment for a digital good without a trusted mediator," in *2019 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, IEEE, 262–267.

Athey, S. and I. Segal (2013): "An efficient dynamic mechanism," *Econometrica*, 81, 2463–2485.

Bogner, A., M. Chanson, and A. Meeuw (2016): "A decentralised sharing app running a smart contract on the ethereum blockchain," in *Proceedings of the 6th International Conference on the Internet of Things*, ACM, 177–178.

Budish, E. (2018): "The economic limits of bitcoin and the blockchain," Working paper.

Carlsson, H. and E. Van Damme (1993): "Global games and equilibrium selection," *Econometrica*, 989–1018.

Cong, L. W. and Z. He (2019): "Blockchain disruption and smart contracts," *The Review of Financial Studies*, 32, 1754–1797.

Ellis, S., A. Juels, and S. Nazarov (2017): "ChainLink: A decentralized oracle network," The white paper of ChainLink.

Galal, H. S. and A. M. Youssef (2019): "Verifiable sealed-bid auction on the Ethereum blockchain," in *Financial Cryptography and Data Security*, ed. by A. Zohar, I. Eyal, V. Teague, J. Clark, A. Bracciali, F. Pintore, and M. Sala, Berlin, Heidelberg: Springer Berlin Heidelberg, 265–278.

Gneezy, U. (2005): "Deception: The role of consequences," *American Economic Review*, 95, 384–394.

Goldreich, O. (2007): *Foundations of cryptography: volume 1, basic tools*, Cambridge university press.

Hörner, J., S. Takahashi, and N. Vieille (2015): "Truthful equilibria in dynamic Bayesian games," *Econometrica*, 83, 1795–1848.

Huberman, G., J. Leshno, and C. Moallemi (2017): "An economic analysis of the bitcoin payment system," Working paper.

——— (2019): "An economist's perspective on the Bitcoin payment system," *American Economic Association Papers and Proceedings*, 109, 93–96.

Jackson, M. O. (1992): "Implementation in undominated strategies: A look at bounded mechanisms," *The Review of Economic Studies*, 59, 757–775.

——— (2001): "A crash course in implementation theory," *Social choice and welfare*, 18, 655–708.

Kosba, A., A. Miller, E. Shi, Z. Wen, and C. Papamanthou (2016): "Hawk: The blockchain model of cryptography and privacy-preserving smart contracts," in *2016 IEEE symposium on security and privacy (SP)*, IEEE, 839–858.

Kreps, D. M., P. Milgrom, J. Roberts, and R. Wilson (1982): "Rational cooperation in the finitely repeated prisoners' dilemma," *Journal of Economic theory*, 27, 245–252.

Krishna, V. (2009): *Auction theory*, Academic press.

Maskin, E. (1999): "Nash equilibrium and welfare optimality," *The Review of Economic Studies*, 66, 23–38.

Maskin, E. and T. Sjöström (2002): "Implementation theory," *Handbook of social Choice and Welfare*, 1, 237–288.

Matsushima, H. (1988): "A new approach to the implementation problem," *Journal of Economic Theory*, 45, 128–144.

——— (2008): "Role of honesty in full implementation," *Journal of Economic Theory*, 139, 353–359.

——— (2019): "Blockchain disables real-world governance," Discussion Paper, CARF-F-459, University of Tokyo.

McAfee, R. P. and J. McMillan (1987): "Auctions and bidding," *Journal of Economic Literature*, 25, 699–738.

——— (1992): "Bidding rings," *American Economic Review*, 579–599.

McCorry, P., S. F. Shahandashti, and F. Hao (2017): "A smart contract for boardroom voting with maximum voter privacy," in *International Conference on Financial Cryptography and Data Security*, Springer, 357–375.

Moore, J. and R. Repullo (1988): "Subgame perfect implementation," *Econometrica*, 1191–1220.

Morris, S. and H. S. Shin (1998): "Unique equilibrium in a model of self-fulfilling currency attacks," *American Economic Review*, 587–597.

Nakamoto, S. (2008): "Bitcoin: A peer-to-peer electronic cash system," The white paper of Bitcoin.

Palfrey, T. R. and S. Srivastava (1991): "Nash implementation using undominated strategies," *Econometrica*, 479–501.

Peterson, J., J. Krug, M. Zoltu, A. K. Williams, and S. Alexander (2015): "Augur: a decentralized oracle and prediction market platform," ArXiv preprint, arXiv:1501.01042.

Postlewaite, A. and X. Vives (1987): "Bank runs as an equilibrium phenomenon," *Journal of Political Economy*, 95, 485–491.

Szabo, N. (1994): "Smart contracts," http://www.fon.hum.uva.nl/rob/Courses/InformationInSpeech/CDROM/Literature/LOTwinterschool2006/szabo.best.vwh.net/smart.contracts.html. Accessed on January 15th 2020.

Tirole, J. (1992): "Collusion and the theory of organizations," in *Advances in Economic Theory: Sixth World Congress, Vol. II*, ed. by J.-J. Laffont, Oxford: Cambridge University Press, 151–206.

Zyskind, G., O. Nathan, and A. Pentland (2015): "Enigma: Decentralized computation platform with guaranteed privacy," ArXiv preprint, arXiv:1506.03471.