June 16, 2020
10:25 ~ 12:10 (remote seminar), University of Tokyo

Mechanism Design with Blockchain Enforcement

Hitoshi Matsushima
University of Tokyo

Shunya Noda
University of British Columbia

1. Introduction

Enforcement of Real-World Business Agreement

Agreed Action Profile $\alpha = (\alpha_1, ..., \alpha_n)$,
Work hard, pay money, ...

Actual Action Profile $a = (a_1, ..., a_n)$
Skipped work, pay nothing, ...

Player i **is innocent** $a_i = \alpha_i$: $\omega_i = 0$

Player i **is guilty** $a_i \neq \alpha_i$: $\omega_i = 1$

How can we penalize (only) guilty players?

Legal Enforcement: Authorized and Trusted Court

Business Agreement

Agreed Action Profile

$$\alpha = (\alpha_1, ..., \alpha_n)$$

Actual Action Profile

$$a = (a_1, ..., a_n)$$

Player i is innocent $a_i = \alpha_i$

Player i is guilty $a_i \neq \alpha_i$

Authorized and Trusted Court

Trial i

Costly Verification to Third Parties
Mandatory Penalties
Privacy Infringement
Elimination of Illegal Activities

We proposed a new method of business enforcement: Blockchain Enforcement

Verification → Incentives

Matsushima (May 2019):

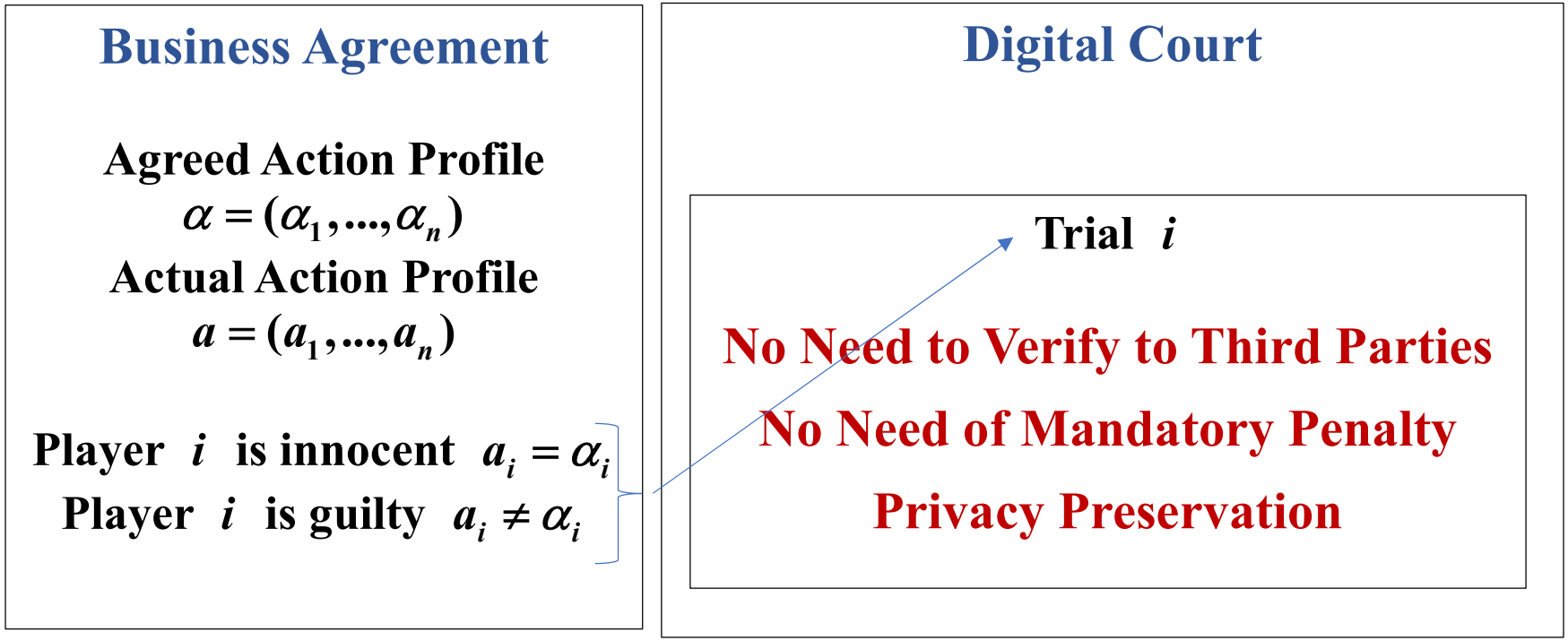
"Blockchain Disables Real-World Governance"
CARF-F-459, U-Tokyo.
Partial Implementation, Cartelization

Matsushima and Noda (2020):

"Mechanism Design with Blockchain Enforcement"
CARF-F-474, U-Tokyo (First Version)
Unique Implementation, Behavioral Theory

Blockchain Enforcement: Digital Court

(No Third Party, Self-Enforcement by Untrusted, Startups)



What is Blockchain?

- **Blockchain is a decentralized ledger technology.**
- **Blockchain works as an electronic payment system.**
- **Blockchain can manage ownership and transaction data in cryptocurrency in a tamper-proof (never-rewrite, safe) manner.**
- **Blockchain can add cryptocurrency its own value as a means of exchange.**
- **Traditional electronic payment (ex. Credit Card):**
  Record keeper is centralized and trusted,
  backing a record to fiat money.
- **Blockchain electronic payment:**
  Record keepers are decentralized and even untrusted,
  not backing a record (cryptocurrency) to fiat money.
- **Importantly, cryptocurrency is programmable on blockchain such as Ethereum.**

Smart Contract

- **Cryptocurrency is programmable.**
  → **Blockchain can allow users to write input-contingent transfer rule as "Smart Contract".**

- **Smart Contract is tamper-proof. Many steps of execution are automated.**
  → **Smart contract can be used as commitment device.**

Smart contract has limitations:

- **Only cryptocurrency transfers are available in automation.**
- **All transactions on blockchain are publicly observable.**
  → **Parties (players) should (can) not input details of their business.**
- **Parties must input by themselves.**
  → **Oracle Problem: We need incentive design in smart contract.**

Players deploy Smart Contract as commitment device.

Smart Contract

$$q = (q_i)_{i \in N},\ q_i : M \to (-\infty, T_i)$$

Players deploy q to blockchain.

→ Each player i deposits T_i in cryptocurrency.

→ They input messages $m = (m_i)$.

→ Each player i automatically pays or burns $q_i(m)$.

Use smart contracts as Self-Enforcement Device.

Digital Court

Trial i

$q^i = (q^i_j)_{j \in N}$, $q^i_j : M^i \to (-\infty, T^i_j)$

Each player $j \in N$ **inputs opinion** $m^i_j \in [0,1]$ **about whether player** i **is innocent** ($\omega_i = 0$) **or guilty** ($\omega_i = 1$)

Each player j **automatically pays or burns** $q^i_j(m^i)$.

Sentence $s^i = s^i(m^i) \in \{0,1\}$

Consider design of digital court strategically to incentivize players to input correctly.

Advantages of Our "Digital Court" Proposal

1. **Digital court needs no verification process to third party:**
   We can save judicial expenses.

2. **Digital court can preserve privacy:**
   Smart contract is publicly observable but includes no detail.

3. **Digital court can make any real-world agreement self-enforcing:**
   Parties can be untrusted, have no long-term relationship

4. **Digital court needs no authority or monopolistic power of control:**
   We use programable money as automated tamper-proof device.

5. **Our design of digital court is simple:**
   Low commissions.
   It can be implemented with current technology.

2. Model

Real-World (Business) Agreement

Agreed Action Profile $\alpha = (\alpha_1, ..., \alpha_n)$

Actual Action Profile $a = (a_1, ..., a_n)$

Actual action profile is observable with each other but not to third parties.

Player i **is innocent** $a_i = \alpha_i$**:** $\omega_i = 0$

Player i **is guilty** $a_i \neq \alpha_i$**:** $\omega_i = 1$

Irrespective of $\omega_i \in \{0,1\}$**, each player** $i \in N$ **becomes a defendant, and has a separate trial on blockchain.**
All players take part in every trial as jurors.

Trial $i \in N$

All players deploy smart contract as transfer rule profile $q^i = (q^i_j)_{j \in N}$, where

$$q^i_j : M^i \to R, \quad M^i = \underset{j \in N}{\times} M^i_j, \quad \sum_{j \in N} q^i_j(m^i) \geq 0.$$

Each player $j \in N$ deposits $T^i_j \equiv \max_{m^i} q^i_j(m^i) \geq 0$ in cryptocurrency.

After all players selecting actual actions a offline, each player j inputs message $m^i_j \in M^i_j \subseteq [0,1]$ about whether defendant i is innocent or guilty.

Each player j automatically pays or burns $q^i_j(m^i)$.

Because of separate trials, we focus on 'Trial 1'.

(Defendant is player 1. Omit subscript and superscript.)

Design I: $M_i = \{0,1\}$ for each $i \in N$

Juror i Inputs either "innocent $m_i = 0$" or "guilty $m_i = 1$".

$$q_1(m) = T_1 s(m_{-1}) + \frac{\eta}{n-1}\sum_{k\neq 1}(m_1 - m_k)^2$$

$$q_i(m) = \frac{\eta}{n-1}\sum_{k\neq i}(m_i - m_k)^2 \quad \text{for all } i \neq 1,$$

where $s : M \to \{0,1\}$ is a sentence function specified as the majority rule:

$$s(m_{-1}) = 1 \qquad \text{if } \sum_{i\neq 1} m_i > \frac{n-1}{2}$$

$$s(m_{-1}) = 0 \qquad \text{otherwise.}$$

η is positive but just tiny.

- Defendant is fined a large amount T_1 if she is convicted ($s(m_{-1}) = 1$).
- Each juror i is incentivized to make her message close to the others' messages according to scoring rule $(m_i - m_k)^2$.

Mis-Coordination:

Player j is rational (purely self-interested), i.e., minimizes $q_j(m)$ in expectation.
Then, in Design I, both $m=(0,...,0)$ and $m=(1,...,1)$ are NE irrespective of $\omega \in \{0,1\}$.
Both truth telling and lying are NE in Design I.

We can generalize this impossibility:

Theorem 1: Suppose all players are rational. Then, irrespective of design of digital court, the set of all Nash equilibria is independent of whether the defendant is innocent or guilty.

If all players are rational,
full (unique) implementation is impossible.

Why?

Only transfer can be automated.
(Any juror always prefers greater transfer.)
No digital-data-based public monitoring is available.

Hence, we cannot use any rationality-based uniqueness devices such as
VCG, Abreu-Matsushima,
Equilibrium Refinements,
Global Games, ……

We need a good explanation about
which equilibrium behavior rational players actually take and why so.

More specifically,
we need a good explanation about
when and why a rational player expects the other players
to behave honestly.

However, rationality-based theories mostly fail.

∴ Let's consider
not only rational (pure self-interest) motive but also
behavioral motives

3. Behavioral Model

Incorporate behavioral aspects into mechanism design theory

Related Literatures:

Reputation Theory:	**Crazy Types** **Gang of Four (1982)**
This paper:	**no reputation, no social pressure**
Behavioral Mechanism Design:	**Preference for Honesty** **Matsushima (2003)**
This paper:	**Whether a player is behavioral (rational, pro-social, or anti-social) is unknown to the others.**

Consider continuum message space $M_i = [0,1]$ instead of $M_i = \{0,1\}$.

Each player i announces $m_i \in [0,1]$ about how likely defendant is to be guilty.

We assume that each player's type is:

Rational

Honest (pro-social motive)

or

Adversarial (anti-social motive).

Rational (R) **minimize monetary payment $q_i(m)$ (in expectation).**

Type R only considers financial gain.

Honest (H) **minimize (in expectation):**

$$\lambda_{i,H} q_i(m) + \left\{ \omega c^1_{i,H}(m_i) + (1-\omega) c^0_{i,H}(m_i) \right\}$$

where $(c^1_{i,H})' < 0,\ (c^1_{i,H})'' > 0,\ (c^0_{i,H})' > 0,\ (c^0_{i,H})'' > 0$

Type H considers both financial gain and psychological cost (pro-social motive).
Type H prefers honest input and hates a big lie (convexity of psychological cost function).

An Extreme case: $\lambda_{i,H} = 0$, **ignore financial gains**
"announce $m_i = \omega$**"**

Adversarial (A) **minimize (in expectation)**

$$\lambda_{i,A} q_i(m) + \left\{ \omega c^1_{i,A}(m_i) + (1-\omega) c^0_{i,A}(m_i) \right\}$$

where $(c^1_{i,A})' > 0,\ (c^1_{i,A})'' > 0,\ (c^0_{i,A})' < 0,\ (c^0_{i,A})'' > 0$

Type A considers both financial gain and psychological cost (anti-social motive).
Type A prefers dishonest input and hates an idiot honesty (convexity of psychological cost function).

An Extreme case: $\lambda_{i,A} = 0$**, ignore financial gains**
"announce $m_i = 1 - \omega$**"**

Remark: **We can envision more diverse behavioral types such as**

always announce "innocent 0"
always announce "guilty 1"
heterogeneity of honest types
heterogeneity of adversarial types
…….

We can generalize our analysis on this line without substantial changes.

Important Features of our behavioral model:

A behavioral motive is state-contingent.
A behavioral type sticks to a specific pattern such as 'be honest' and 'be adversarial'.

Incomplete Information:

The others do not know which type each player is.

Player i is	**honest (H)**	**with prob.**	$\delta_{i,H}$
	adversarial (A)	**with prob.**	$\delta_{i,A}$
	rational (R)	**with prob.**	$1-\delta_{i,H}-\delta_{i,A}$

Both $\delta_{i,H}$ and $\delta_{i,A}$ are just tiny.

We slightly modify Design I:

Design II: Modify Design I by replacing $M_i = \{0,1\}$ with continuum message space:

$$M_i = [0,1] \text{ for each } i \in N.$$

We have a possibility result: Extreme Case

Suppose all players are never adversarial, and a single player could be honest:

$$\delta_{i,A} = 0 \quad \text{for all} \quad i \in N,$$

$$\delta_{i,H} > 0 \quad \text{for some} \quad i \in N.$$

Then, we have a possibility result:

Rational types input full honesty ($m_i^{\omega}(R) = \omega$ for all $i \in N$) as unique BNE behavior.

Why? **Any rational player (except i) attempts to announce more honestly than the average of the other rational players, because of i's potential honesty.**

→ **Tail-chasing competition reaches full honesty.**

Theorem 2 (Uniqueness in General):

Suppose

$$\delta_{i,H} + \delta_{i,A} > 0 \text{ for some } i \in N.$$

Then, irrespective of $\omega \in \{0,1\}$, there exists unique BNE, m^{ω}, as unique iteratively undominated strategy profile (dominance solvable), where we denote

$$m^{\omega} = (m_i^{\omega})$$
$$m_i^{\omega} = (m_i^{\omega}(R), m_i^{\omega}(H), m_i^{\omega}(A)) \in [0,1]^3.$$

Proof of Theorem 2 depends on:

- **Behavioral types are less elastic than rational type (stick to patterns)**
  - → **BNE is expressed by a fix point of some contractive mapping**
  - → **Uniqueness of fixed point (Edelstein's Fix Point Theorem, 62)**

- **Continuum of message spaces and continuity of preferences:**
  - → **Existence of fixed point**

- **Supermodular game: Convexity of psychological cost, proper scoring rule**
  - → **Uniqueness of BNE implies dominance solvable.**

Properties of Unique BNE

* $\left|\omega - m_i^{\omega}(X)\right|$ **is decreasing in** $\delta_{j,H}$ **and increasing in** $\delta_{j,A}$ **for all** $i \in N$, $j \neq i$, **and** $X \in \{R, H, A\}$.

  ∴ **The more likely the other players are to be honest (adversarial), the more (less) honest a rational player behaves.**

* **Turning over A and H, we have** $1 - m^{\omega}$ **instead of** m^{ω}.

  ∴ **We have symmetry between A and H.**

From these properties, we can say:

- **Whenever players are more likely to be honest type than adversarial type, correct judgement is supported by unique BNE.**

- **Whenever players are less likely to be honest type than adversarial type, incorrect judgement is supported by unique BNE.**

More on Extreme Case:

($\lambda_{i,H} = \lambda_{i,A} = 0$, i.e., behavioral types never consider financial gains)

In the extreme case, we can calculate unique BNE as reduced forms:

$$m_i^{\omega}(H) = \omega, \;\; m_i^{\omega}(A) = 1 - \omega,$$

$$m_i^1(R) = \frac{\sum_{j \in N} m_j^1(R) - \delta_{i,H}}{n - \delta_{i,H} - \delta_{i,A}}$$

$$m_i^0(R) = 1 - \frac{\sum_{j \in N} \{1 - m_j^0(R)\} - \delta_{i,H}}{n - \delta_{i,H} - \delta_{i,A}}, \text{ where}$$

$$\sum_{j \in N} m_j^1(R) = \sum_{j \in N} \{1 - m_j^0(R)\} = \frac{\sum_{j \in N} \frac{\delta_{j,H}}{n - \delta_{j,H} - \delta_{j,A}}}{\sum_{j \in N} \frac{1}{n - \delta_{j,H} - \delta_{j,A}} - 1}.$$

Example 1 (Neutrality): **Assume**

$$\delta_{i,H} = \delta_H \quad \text{and} \quad \delta_{i,A} = \delta_A \quad \text{for all} \quad i \in N.$$

We have

$$m_i^1(R) = \frac{\delta_H}{\delta_H + \delta_A} \quad \text{and} \quad m_i^0(R) = \frac{\delta_A}{\delta_H + \delta_A}.$$

If $\delta_H > \delta_A$, unique BNE yields correct judgement.
If $\delta_H < \delta_A$, unique BNE yields incorrect judgement.

Example 2: **Assume each player is either potential honesty or potential liar, that is,**

$$\max[\delta_{i,H},\delta_{i,A}]=\delta \quad \text{and} \quad \min[\delta_{i,H},\delta_{i,A}]=0.$$

We have

$$m_i^1(R)=\frac{\tilde{n}-\delta}{n-\delta} \quad \text{and} \quad m_i^0(R)=1-\frac{\tilde{n}-\delta}{n-\delta} \quad \text{for} \quad \tilde{n} \quad \text{players,}$$

$$m_i^1(R)=\frac{\tilde{n}}{n-\delta} \quad \text{and} \quad m_i^0(R)=1-\frac{\tilde{n}}{n-\delta} \quad \text{for} \quad n-\tilde{n} \quad \text{players.}$$

With $\tilde{n}\geq 1/2$ (potential honesty is majority), unique BNE yields correct judgement.

With $\tilde{n}<1/2$ (potential liar is majority), unique BNE yields incorrect judgement.

State-Contingent Belief on $(\delta_{i,H}, \delta_{i,A})$:

In Example 2, suppose that

$$\text{Player } i \text{ is innocent } (a_i = \alpha_i) \rightarrow \delta_{i,H} = \delta \text{ and } \delta_{i,A} = 0$$
$$\text{Player } i \text{ is guilty } (a_i \neq \alpha_i) \rightarrow \delta_{i,H} = 0 \text{ and } \delta_{i,A} = \delta$$

Then, if more than half are guilty, (not guilty but) innocent is penalized.

Alternative Interpretation:

Player i is guilty, but not on purpose.
$\rightarrow$ The other players still believe $\delta_{i,H} = \delta$ and $\delta_{i,A} = 0$.

4. Legal Purposes, Illegal Purposes, Logistics

Illegal Purposes with Logistics:

Illegal Drug is an example.

Prevent drug crime by observing logistics.

Illegal Purpose without Logistics:

It is hard to crack down illegal behavior without logistics.

Serious tension between privacy and illegal activities.

Cartelization is an example.

Purpose-Contingent Beliefs on $(\delta_{i,H}, \delta_{i,A})$:

Fear of adversarial type relieves sound business from illegal cartels.

Optimistic View:

Business is legal $\rightarrow$ $\delta_H > \delta_A$

Business is illegal (Cartelization) $\rightarrow$ $\delta_H < \delta_A$

Pessimistic View:

Even illegal business satisfies $\delta_H > \delta_A$.

$\rightarrow$ **Any business chance (with three or more players) fails due to fear of blockchain-based Cartelization.**

$\rightarrow$ **Regulator may have to ban blockchain use.**

How $(\delta_{i,H}, \delta_{i,A})$ is determined off equilibrium path?

"Social Image" may matter.
(but no social pressure, no reputation)

I have a better social image.
→ I expect you play more honestly.
→ I prefer playing more honestly.
→ I expect you play more honestly.
→ …….

5. False Charge Problem

Innocent defendant should not be fined a large amount.
Honest-type juror should not be fined a large amount.

However, in Design II, with adversarial type ($\delta_{i,A} > 0$), innocent defendant is fined a large amount with a positive probability.

$\therefore$ We need alternative design of digital court.
(but, maybe more complicated and ad hoc……)

Design II': Each player inputs multiple (Z) messages to a trial. We have Z sub-trials, and have Z sentences, $s(m_{-1}(1)), \ldots, s(m_{-1}(Z))$.

$$m_i = (m_i(1),...,m_i(Z)) \in [0,1]^Z$$

$$q_1(m) = \frac{T_1}{Z}\sum_{z=1}^{Z} s(m_{-1}(z))$$
$$+\frac{\eta(1)}{n-1}\sum_{k\neq 1}\left\{m_1(1)-m_k(1)\right\}^2 + \sum_{z\neq 1}\eta(z)\left\{(m_1(z)-s(m_{-1}(z-1))\right\}^2$$

and for each $i \neq 1$,

$$q_i(m) = \frac{\eta(1)}{n-1}\sum_{k\neq i}\left\{m_i(1)-m_k(1)\right\}^2 + \sum_{z\neq 1}\eta(z)\left\{(m_i(z)-s(m_{-1}(z-1))\right\}^2$$

Incentive in 1-st Input:

Scoring Rule:

$$\left\{m_i(1) - m_k(1)\right\}^2$$

Incentive in z-th Input:

Punishment rule based on distance from $(z-1)$-th sentence $s(m_{-1}(z-1))$:

$$\left\{(m_i(z) - s(m_{-1}(z-1))\right\}^2$$

We have Z sentences:

Each sentence punishes defendant by small amount:

$$T_1 / Z.$$

Behavioral Model (modified):

Honest **minimize (in expectation):**

$$\lambda_{i,H} q_i(m) + \sum_{z=1}^{Z} \lambda_{i,H}(z) \left\{ \omega c_{i,H}^{1}(m_i(z)) + (1-\omega) c_{i,H}^{2}(m_i(z)) \right\}$$

Adversarial **minimize (in expectation):**

$$\lambda_{i,A} q_i(m) + \sum_{z=1}^{Z} \lambda_{i,A}(z) \left\{ \omega c_{i,A}^{1}(m_i(z)) + (1-\omega) c_{i,A}^{2}(m_i(z)) \right\}$$

where $\sum_{z=1}^{Z} \lambda_{i,H}(z) = \sum_{z=1}^{Z} \lambda_{i,A}(z) = 1$ **and** $\lambda_{i,A} > 0$.

Assume in Design II' that

$$M_i(1) = [0,1],$$
$$M_i(z) = \{0,1\} \text{ for all } z \neq 1,$$

and

$$\eta(z) > \lambda_{i,A}(z)\left\{c^0_{i,A}(0) - c^0_{i,A}(1)\right\} \text{ for all } i \in N \text{ and } z.$$

With this assumption, Design II' is dominance solvable.
In unique BNE, A rational player inputs correctly from 2 to Z:

$$m_i^{\omega}(z;R) = m_i^{\omega}(z;H) = m_i^{\omega}(z;A) = \omega \text{ for all } z \neq 1.$$

Consider a sufficiently large Z.
Assume in modified behavioral model that

both $\lambda_{i,A}(1)$ and $\lambda_{i,A}(2)$ are sufficiently small compared with $\lambda_{i,A}$.

With this assumption, we can set $\eta(1)+\eta(2)$ close to zero.

With these assumptions, Design II' solves false charge problem:

- **Innocent defendant is never fined more than a small amount T_1/Z.**
- **Honest juror is never fined more than a small amount $\eta(1)+\eta(2)$.**

6. Further Remarks

6.1. Coalition in Digital Court

A guilty defendant asks you:
"Please vote for innocence. I will give you $100."

Do you accept this request?

- **If you are trusted third party,** **"No"**
- **If you are untrusted third party,** **"Yes"**
- **If you are victim (Digital Court),** **"No, maybe"**
  Reciprocal Retaliation

6.2. Deposit Savings

A player may have to deposit a large amount in advance:
cf. Auction: Bidders deposit, before or after a win?

Can we save deposit?

→ **Consider dynamics with sequential business opportunities.**

- **We can reuse a deposit for many businesses (Matsushima, 2012)**
- **We can add deposit little by little (a la reputation-building).**

7. Conclusion

- **We demonstrated blockchain enforcement, a new method.**
- **We introduced digital court as a commitment device in cryptocurrency. We characterized the case that a digital court makes business agreement self-enforcing.**
- **By replacing legal enforcement with blockchain enforcement, we can eliminate verification processes, saving judicial expenses and preserving privacy.**
- **This method, however, can be used in illegal applications such as cartelization. In the worst case, blockchain plucks the bud of all business opportunities.**
- **To understand appropriate policies, it is important as future research to develop systematic analysis in theory, experiment, and social implementation by unifying real-world with virtual-world.**