AI Needs Blockchain: Trustless Solutions to Failures in Machine to Colloidal Markets

John P. Conley

Vanderbilt University

April 19, 2024

2024 NSF/CEME Decentralization Conference, Vanderbilt University

Introduction

Creating value through exchange is often sequential rather than atomic.

One party commits resources or assets in exchange for the promise of compensation or a share of the resulting payoff in the future.

What prevents the counterparty from absconding with the proceeds in this type of sequential trust game?

In the one-shot case, nothing at all. Markets fail, and gains from trade are not created.

Introduction

In the repeated game, however, a Folk Theorem may apply.

Societies have long had informal mechanisms that allow their members to extend trust to one another on the basis of reputation.

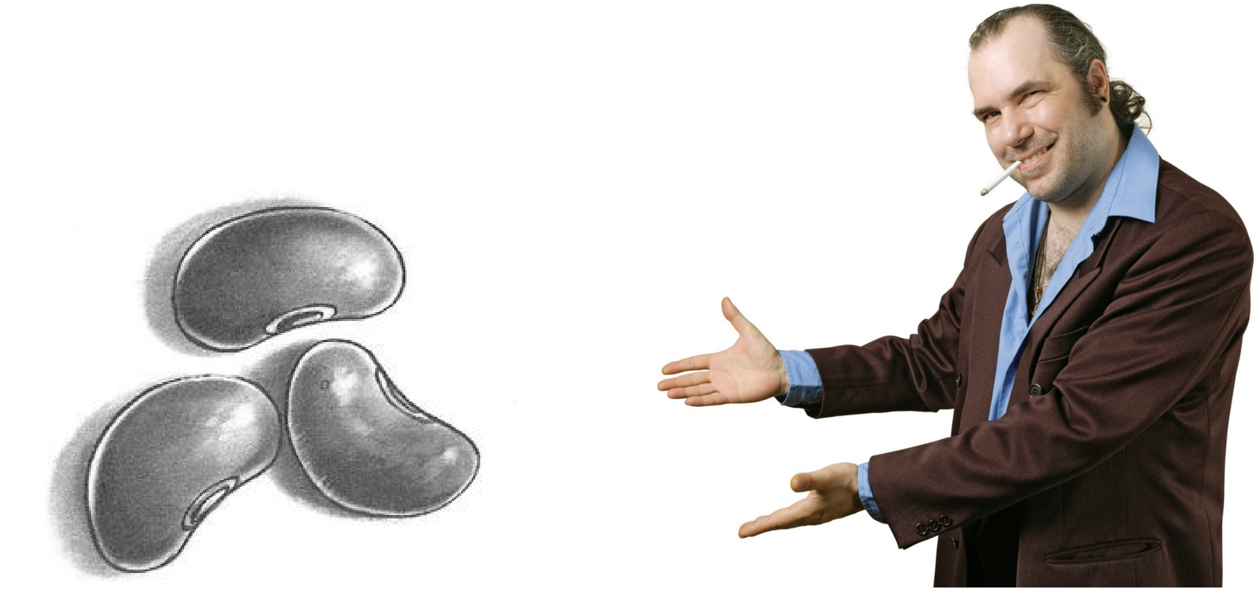
Bad behavior is punished by a loss of reputation and being excluded from beneficial exchanges in the future.

Such mechanism rely on several elements:

- Identification (non-anonymity).
- Establishment and dissemination of credible histories of behavior.
- An expectation of future interactions.

Don’t Trust Nobody

Extending trust to unknown agents, agents without reputations to lose, or that you will likely never see again, is a bad idea.



Can I interest you in some magic beans?

Large Economies

Identifying agents, and knowing their histories becomes more difficult in larger economies.

We have established elaborate systems of identity, attestors, credit rating, background checks, and so on, to facilitate the extenuation of trust between people in modern societies.

These institutions are mainly designed for biological persons, although they can be extended to legal persons, especially those that have a physical existence.

Virtual Economies

These institutions of identity and reputation, on the other hand, do not extend well to virtual and electronic interactions between humans.

How they might extend to Artificial Agents is even less clear.

If we can't fix this, then we may not be able to realize gains from trade in virtual environment.

In particular, Machine to Colloidal markets (M2C?) may fail.

Turing Tests and Identity

"On the Internet, nobody knows you're a dog."

If an AI can pass a **Turing Test** (it only has to get a D−), there is no way in a virtual environment to tell that it is non-colloidal.

Questions

- Do AIs even have an identity?
- Do AIs have preferences or objectives?
- Do AIs have a sense of individuality or continuity?
- Even if AIs they have identity and preferences, do they care about the welfare of a future "self"?
- Are AIs "rational" in an economic sense?
- Can we extend our mechanisms to Intelligent Machines?

Answers

- I don’t know, but I don’t think it matters.
- I have no idea.
- I have even less of an idea.
- Your guess is as good as mine.
- See above .
- Yes.

One-Shot Game

We consider a game with two types of anonymous agents: Biological Humans and Machine Intelligences, which we call **Biologicals** and **Mechanicals:**

$$\text{Biologicals}: \quad b \in \{1, \ldots B\} \equiv \mathcal{B}$$

$$\text{Mechanicals}: \quad m \in \{1, \ldots m\} \equiv \mathcal{M}.$$

Processes

Mechanicals have a comparative advantage at executing certain types of tasks which we call **Processes:**

$$\mathbf{Proc}: \mathrm{INPUT} \Rightarrow \mathrm{OUTPUT}$$

where

$$\mathrm{input}_{\mathrm{i}} \in \{\mathrm{input}_1, \ldots \mathrm{input}_{\mathrm{I}}\} \equiv \mathrm{INPUT}$$

$$\mathrm{output}_{\mathrm{o}} \in \{\mathrm{output}_1, \ldots \mathrm{output}_{\mathrm{O}}\} \equiv \mathrm{OUTPUT}$$

$$\mathrm{p} \in \{1, \ldots \mathrm{P}\} \equiv \mathcal{P}, \mathrm{i} \in \{1, \ldots \mathrm{I}\} \equiv \mathcal{I}, \in \{1, \ldots \mathrm{O}\} \equiv \mathcal{O}.$$

Cost of Executing a Process correctly to a Mechanical:

$$\mathbf{CostProc}: \mathrm{PROC} \Rightarrow (0, \overline{\mathrm{CP}}] \text{ where } \mathrm{CP} \in (0, \overline{\mathrm{CP}}]$$

Audits and Verifiers

Audits that confirm that a Mechanical has executed a process correctly are conducted by external agents called **Verifiers.**

$$\textbf{Verify}: \text{PROC} \times \text{INPUT} \times \text{OUTPUT} \Rightarrow \{\text{CORRECT}, \text{MALICIOUS}\}$$

such that $\forall\, p \in \mathcal{P},\ i \in \mathcal{I},$ and $o \in \mathcal{O}$

$$\text{Verify}(\text{Proc}_p, \text{input}_i, \text{output}_o) = \text{CORRECT} \Leftrightarrow \text{Proc}_p(\text{input}_i) = \text{output}_o$$

$$\text{Verify}(\text{Proc}_p, \text{input}_i, \text{output}_o) = \text{MALICIOUS} \Leftrightarrow \text{Proc}_p(\text{input}_i) \neq \text{output}_o$$

Cost of Verifying an Execution of a Process to a Verifier:

$$\textbf{CostVerify}: \text{PROC} \Rightarrow (0, \overline{\text{CV}}]\ \textbf{where}\ \text{CV} \in (0, \overline{\text{CV}}]$$

Trust Games – The Biological

Biologicals and Mechanicals play a sequential **Trust Game**.

Biologicals move first and choose either to make an **Offer** or **PASS**.

An offer consists of a **Fee** paid in advance to Mechanicals to compensate them for executing a process:

$$\text{Fee} \in [0, \overline{\text{F}}]$$

and **p**, an **Audit Probability**:

$$\text{p} \in [0, 1]$$

If a Biological decides to PASS, no fees or inputs are sent.

Trust Games – The Mechanical

Mechanical moves second after seeing the Biological's action.

- If the Biological makes an offer, the Mechanical decides whether to accept or reject it.
- If he accepts, the Biological sends the offered fee and his input to The Mechanical, and $(p \times CV)$ to a Verifier.
- The Mechanical then chooses **CORRECT** or **MALICIOUS**, execution, and sends an output to the Biological.
- Alternatively, the Mechanical can decline the offer and choose **NULL** execution.
- In this case, the game is over, and no fees, inputs, or outputs are exchanged. If the Biological chooses to PASS, then NULL execution is the only action available to the Mechanical.

Actions

Formally, the **Action Space** is defined as follows:

$$a^{b} \in \{(\mathrm{Fee}, p) \in [0, \bar{F}]\times[0, 1], \mathrm{PASS}\} \equiv \mathcal{A}^{b}$$

$$a^{m} \in \{\mathrm{CORRECT}, \mathrm{MALICIOUS}, \mathrm{NULL}\} \equiv \mathcal{A}^{m}$$

Biological Objectives

The one-period **Utility Function of Biologicals** if an offer is accepted depends on how it is executed:

$$\mathbf{Utility}^{\mathbf{b}}\colon \mathrm{PROC} \times \mathrm{INPUT} \times \mathrm{OUTPUT} \Rightarrow [0, \overline{\mathrm{U}}]$$

where if

$$\mathrm{Verify}(\mathrm{Proc}_p, \mathrm{input}_i, \mathrm{output}_o) = \mathrm{MALICIOUS},$$

then

$$\mathrm{Utility}^{b}(\mathrm{Proc}_p, \mathrm{input}_i, \mathrm{output}_o) = 0.$$

We assume that Biologicals cannot determine if a process was executed correctly unless they explicitly verify it. Further, we assume that Biologicals are unable to attribute any increase or decrease in their utility to how a Mechanical chooses to execute a given process

Mechanical Objectives

We assume that Mechanicals value fees net of processing costs.

- This might be explained by an existence of an unmodeled Biological agent who instantiates a given Mechanical, programs its behavior, and receives any net value generated by his creation.
- It might also reflect the need of an autonomous Mechanical for resources to exist or replicate.
- We also allow for the possibility that MALICIOUS execution might give Mechanical a higher payoff, all else equal:

Net Value of Malicious Execution to a Mechanical:

$$\mathbf{MaliciousValue}: \mathrm{INPUT} \Rightarrow (0, \overline{\mathrm{MV}}] \text{ where } \mathrm{MV} \in (0, \overline{\mathrm{MV}}]$$

Payoffs

Given some $(\mathrm{Proc}_p, \mathrm{input}_i) \in \mathrm{PROC} \times \mathrm{INPUT}$, the **Payoff Functions** for agents are defined as follows:

$$F: \mathcal{A}^b \times \mathcal{A}^m \Rightarrow \mathbb{R}^2 \equiv (F^b(a^b, a^m), F^m(a^b, a^m))$$

where $\forall\ (\mathrm{Fee}, p) \in [0, \bar{F}] \times [0, 1]$,

$$F^b((\mathrm{Fee}, p), \mathrm{CORRECT}) =$$

$$\mathrm{Utility}^b(\mathrm{Proc}_p, \mathrm{input}_i, \mathrm{Proc}_p(\mathrm{input}_i)) - \mathrm{Fee} - p \times \mathrm{CostVerify}(\mathrm{Proc}_p)$$

$$F^b((\mathrm{Fee}, p), \mathrm{MALICIOUS}) = - \mathrm{Fee} - p \times \mathrm{CostVerify}(\mathrm{Proc}_p) - \varepsilon$$

$$F^b((\mathrm{Fee}, p), \mathrm{NULL}) = 0$$

$$F^b(\mathrm{PASS}, \mathrm{NULL}) = 0$$

and

$$F^m((\mathrm{Fee}, p), \mathrm{CORRECT}) = \mathrm{Fee} - \mathrm{CostProc}(\mathrm{Proc}_p)$$

$$F^m((\mathrm{Fee}, p), \mathrm{MALICIOUS}) = \mathrm{Fee} + \mathrm{MaliciousValue}(\mathrm{input}_i)$$

$$F^m((\mathrm{Fee}, p), \mathrm{NULL}) = 0$$

$$F^m(\mathrm{PASS}, \mathrm{NULL}) = 0$$

The Two-Player One-Shot Game

A **Strategy for a Biological** is a choice from his action space.

A **Strategy for a Mechanical** is any mapping from the Biological's action space to CORRECT, MALICIOUS, or NULL execution such that PASS always maps to NULL execution:

$$\mathrm{s}^{\mathrm{b}} \in \mathcal{A}^{\mathrm{b}} \equiv \mathcal{S}^{\mathrm{b}}, \quad \mathrm{s}^{\mathrm{m}} : \mathcal{A}^{\mathrm{b}} \Rightarrow \mathcal{A}^{\mathrm{m}},$$

such that

$$\forall\ \mathrm{s}^{\mathrm{m}} \in \mathcal{S}^{\mathrm{m}}, \mathrm{s}^{\mathrm{m}}(\mathrm{PASS}) = \mathrm{NULL}.$$

A **Strategy Profile** is denoted:

$$\mathrm{S} \equiv (\mathrm{s}^{\mathrm{b}}, \mathrm{s}^{\mathrm{m}}) \in \mathcal{S}^{\mathrm{b}} \times \mathcal{S}^{\mathrm{m}} \equiv \mathcal{S},$$

where $\mathcal{S}^{\mathrm{b}}$ and $\mathcal{S}^{\mathrm{m}}$ denote the **Strategy Spaces** for Biologicals and Mechanicals, respectively.

Equilibrium

Given some $(\mathrm{Proc}_p, \mathrm{input}_i) \in \mathrm{PROC} \times \mathrm{INPUT}$, a strategy profile,

$$S \equiv (s^b, s^m) \in \mathcal{S}$$

is a **Subgame Perfect Equilibrium** (**SPE**) if:

$$\forall\ \bar{s}^b \in \mathcal{S}^b,\ F^b(s^b, s^m(s^b)) \geq F^b(\bar{s}^b, s^m(\bar{s}^b))$$

and

$$\forall\ \bar{s}^b \in \mathcal{S}^b, \forall\ \bar{s}^m \in \mathcal{S}^m,\ F^m(\bar{s}^b, s^m(\bar{s}^b)) \geq F^b(\bar{s}^b, \bar{s}^m(\bar{s}^b)).$$

Note that the Mechanical's strategy must be payoff maximizing for any action the Biological chooses, that is, for every subgame.

Result

Theorem 1: *Given some* $(\text{Proc}_p, \text{input}_i) \in \text{PROC} \times \text{INPUT}$,

$$S = (s^b, s^m) \in \mathcal{S}$$

is an SPE of the one-shot game if and only if:

$$s^b = \text{PASS}$$

$$s^m(\text{PASS}) = \text{NULL}$$

and

$$s^m(\text{Fee}, p) = \text{MALICIOUS}, \forall\ (\text{Fee}, p) \in [0, \bar{F}] \times [0, 1].$$

We see that in the one-shot game Biologicals and Mechanicals are stuck in an SPE that does not allow them to realize the higher payoffs each would receive from reaching an agreement for CORRECT execution.

The Two-Player Repeated Game

Consider the case where one Biological one Mechanical play the sequential game an infinite number of times in succession.

Each agent chooses an action in each period which results in one of four observable **Events** occurring:

COR ≡ Correct: The Biological makes an offer, the Mechanical accepts, and an audit confirms CORRECT execution.

MAL ≡ Malicious: The Biological makes an offer, the Mechanical accepts, and an audit proves MALICIOUS execution.

UNC ≡ Uncertain: The Biological makes an offer, the Mechanical accepts, and no audit takes place.

NUL ≡ Null: The Biological chooses PASS, or the Mechanical chooses NULL.

History

The **Period t History of Play** is the set of events realized up to the end of period t **+** 1.

$$(h_0, \dots h_t) \equiv H_t \in \underbrace{\mathcal{H} \times \dots \times \mathcal{H}}_{\text{t + 1 times}} \equiv \mathcal{H}_t \subset \mathcal{H}_\infty \equiv \mathcal{H} \times \mathcal{H} \times \dots$$

where

$$\forall\ t \in \mathcal{T}\ h_t \in \mathcal{H} \equiv \{\mathrm{COR}, \mathrm{MAL}, \mathrm{UNC}, \mathrm{NUL}\}.$$

A period t history of play in which there have been no successful audits, the Biological has never chosen to PASS, and the Mechanical has never chosen NULL execution, is called a **Cooperative History**:

$$H_t \in \mathcal{H}_\infty^{\text{coop}} \subset \mathcal{H}_\infty \text{ such that } \forall\ t \in \mathcal{T}, h_t \in \{\mathrm{COR}, \mathrm{UNC}\}.$$

Strategies

A **Period t Strategy** for Biologicals is any mapping from period t histories into the Biological action space:

$$\forall\ t \in \mathcal{T},\ s_t^b : \mathcal{H}_t \Rightarrow \mathcal{A}^b \text{ and } s_t^b \in \mathcal{S}_t^b.$$

A **Period t Strategy** for Mechanicals is any mapping from period t histories and the Biological action space into the Mechanical action space such that PASS always maps to NULL execution:

$$\forall\ t \in \mathcal{T},\ s_t^m : \mathcal{H}_t \times \mathcal{A}^b \Rightarrow \mathcal{A}^m$$

such that

$$s_t^m(H_t, \text{PASS}) = \text{NULL and } s_t^m \in \mathcal{S}_t^m.$$

Profiles and Beliefs

A **Strategy Profil**e for the repeated game is denoted:

$$(S^{b}_{\infty},\ S^{m}_{\infty}) \in \mathcal{S}^{b}_{\infty}\times\mathcal{S}^{m}_{\infty} \text{ where } S^{x}_{\infty} \in \prod_{t=0}^{\infty} \mathcal{S}^{x}_{t} \equiv \mathcal{S}^{x}_{\infty}$$

Biologicals only know for certain the history of play up to the current period, t, while the Mechanical knows both this, and the action taken by the Biological.

This constraint is reflected in the arguments that the strategy mappings take. Each must speculate about the actual strategies used their counterparties, and this affects how they evaluate best-responses.

Beliefs

Period t Beliefs about strategies are denoted as follows:

$$\forall\ t \in \mathcal{T}\ \beta_t^m \in \mathcal{S}_t^m \text{ and } \beta_t^b \in \mathcal{S}_t^b.$$

Arbitrary beliefs about complex sequences of strategies for an infinite future are computationally expensive to form and work with, and can rationalize many otherwise implausible equilibrium outcomes.

Consistency requires that agents believe that their counterparties will behave identically in essentially identical situations in all future periods.

The situations in two distinct periods are “essentially identical” if the histories are either both cooperative, or both non-cooperative, and in the case of the Mechanical, the Biological takes the same action.

Consistent Beliefs

Formally, a **Consistent Belief Profile** is defined as follows:

$$(B^b_\infty, B^m_\infty) \in \mathcal{C}^* \mathcal{S}^b_\infty \times \mathcal{C}^* \mathcal{S}^m_\infty \subset \mathcal{S}^b_\infty \times \mathcal{S}^m_\infty$$

is a consistent belief profile if

$$\forall\ t, \bar{t} \in \mathcal{T} \text{ and } \forall\ a^b \in \mathcal{A}^b$$

if

$$H_t, H_{\bar{t}} \in \mathcal{H}^{coop}_\infty,$$

then

$$\beta^b_t(H_t) = \beta^b(H_{\bar{t}}) \text{ and } \beta^m_t(H_t, a^b) = \beta^m_{\bar{t}}(H_{\bar{t}}, a^b)$$

and if

$$H_t, H_{\bar{t}} \notin \mathcal{H}^{coop}_\infty,,$$

then

$$\beta^t_b(H_t) = \beta^b(H_{\bar{t}}) \text{ and } \beta^m_t(H_t, a^b) = \beta^m_{\bar{t}}(H_{\bar{t}}, a^b).$$

Subgames

Subgames for Biologicals start at the beginning of each period $T \in \mathcal{T}$, and are defined by a realized history:

$$H_T \in \mathcal{H}_T.$$

Subgames for Mechanicals start after the Biological has chosen an action, and so depend on both this realized action, and the realized history at the beginning of the period, $(H_T, a_T^b) \in \mathcal{H}_T \times \mathcal{A}^b$.

We assume both Biologicals and Mechanicals discount the future at some rate $\rho \in (0, 1)$ and denote the one period **Discount Factor**:

$$r = (1 - \rho) \in (0, 1).$$

Expected Payoffs

Using this, we denote the **Expected Payoff of a Subgame** defined by H_T for a strategy profile, $(S^b_\infty, S^m_\infty) \in \mathcal{S}^b_\infty \times \mathcal{S}^m_\infty$, as follows:

$$\mathbf{EPO^x}: \mathcal{T} \times \mathcal{H}_T \times \mathcal{S}^b_\infty \times \mathcal{S}^m_\infty \Rightarrow \mathbb{R} = \mathrm{EPO}^x(t, H_t, S^b_\infty, S^m_\infty)$$

Note that $\mathrm{EPO}^x(0, H_0, S^b_\infty, S^m_\infty)$ is the expected payoff to agent x of the supergame.

Value of the Continuation Game

The **Value of the Continuation Game** is the maximum expected payoff to agents when they play the best possible strategy in a period T subgame defined by some history H_t given a fixed strategy for their counterparties:

$$\mathbf{MaxEPO^b}: \mathcal{T} \times \mathcal{H}_\infty \times \mathcal{S}^b_\infty \equiv \underset{\bar{S}^b_\infty \in \mathcal{S}^b_\infty}{\text{Max}} \ \text{EPO}^b(T, H_T, \bar{S}^b_\infty, S^m_\infty).$$

$$\mathbf{MaxEPO^m}: \mathcal{T} \times \mathcal{H}_\infty \times \mathcal{S}^m_\infty \equiv \underset{\bar{S}^m_\infty \in \mathcal{S}^m_\infty}{\text{Max}} \ \text{EPO}^m(T, H_T, S^b_\infty, \bar{S}^m_\infty)$$

Consistent Subgame Perfect Equilibrium

A strategy profile, $(S^b_\infty, S^m_\infty) \in \mathcal{S}^b_\infty \times \mathcal{S}^m_\infty$, is a **Consistent Subgame Perfect Equilibrium (CSPE)** if:

$$\forall\ \bar{S}^b_\infty \in \mathcal{S}^b_\infty, \forall\ \bar{S}^m_\infty \in \mathcal{S}^m_\infty, \forall\ T \in \mathcal{T}, \text{ and } \forall\ H_T \in \mathcal{H}_T$$

$$EPO^b(T, H_T, S^b_\infty, B^m_\infty) \geq EPO^b(T, H_T, \bar{S}^b_\infty, B^m_\infty)$$

$$EPO^m(T, H_T, B^b_\infty, S^m_\infty) \geq EPO^m(T, H_T, B^b_\infty, \bar{S}^m_\infty)$$

where

$$(B^b_\infty, B^m_\infty) \in \mathcal{C}^*\mathcal{S}^b_\infty \times \mathcal{C}^*\mathcal{S}^m_\infty$$

$$\forall\ T \in \mathcal{T},\ \beta^b_T = s^b_T, \forall\ T > 0,\ \beta^m_T = s^m_{(T-1)}$$

and

$$\beta^m_0 \in \mathcal{S}^m_0$$

such that

$$EPO^m(0, H_0, B^b_\infty, B^m_\infty) = \text{MaxEPO}^m(0, H_0, B^b_\infty).$$

Result

Theorem 2: *If*

$$\mathrm{S}^{\mathrm{b}}_{\infty} = \mathrm{Grim}_{\infty} \text{ and } \mathrm{S}^{\mathrm{m}}_{\infty} = \mathrm{MinAccept}_{\infty},$$

then

$$(\mathrm{S}^{\mathrm{b}}_{\infty}, \mathrm{S}^{\mathrm{m}}_{\infty}) \in \mathcal{S}^{\mathrm{b}}_{\infty} \times \mathcal{S}^{\mathrm{m}}_{\infty},$$

is a Consistent Subgame Perfect Equilibrium.

$\mathrm{S}^{\mathrm{b}}_{\infty} = \mathrm{Grim}_{\infty}$ is a grim trigger strategy in which the Biological makes a certain offer each period if and only if the history is cooperative.

$\mathrm{S}^{\mathrm{m}}_{\infty} = \mathrm{MinAccept}_{\infty}$ is a grim trigger strategy in which the Mechanical accepts an offer if and only if the history is cooperative and the expected payoff is above some lower bound.

Discussion

The condition that determines whether the future is cooperative or noncooperative:

$$EPO_C(Fee, p) \equiv \frac{Fee - CP}{(1 - r)} \geq \frac{Fee + MV}{(1 - r + rp)} \equiv EPO_D(Fee, p)$$

satisfies all of our intuitions over fee and audit structure.

- $Fee \geq CP$. That is, fee must always cover the cost of processing. Otherwise, since $Fee + MV > 0$, the inequality could not be satisfied.
- $CP\uparrow$, or $MV\uparrow$, implies either $Fee\uparrow$, or $p\uparrow$. That is, if either the cost of processing, or the value of MALICIOUS execution goes up, then the Biological must either raise the fee offered, or increase the probability of an audit to compensate.

More Discussion

- $p = 1$ implies $(1 - r + rp) = 1$. That is, the payoff from defection is equal to the payoff the Mechanical receives in a single period, since being caught is a certainty if $p = 1$.
- $r \rightarrow 1$ implies $Fee - CP \rightarrow 0$. That is, as agents discount the future less heavily, even small surpluses of fees over processing costs result is high expected payoffs for the Mechanical. On the other hand, $(1 - r + rp) \rightarrow p$. Thus, for fixed, but small probabilities of audit, the relative value of MALICIOUS execution ends up being smaller than the expected value of choosing the CORRECT forever.

Still More Discussion

Also note that the discount rate between periods depends on the length of the period. If a game is played daily, or several times a day, the discount rate gets closer and closer to $r = 1$. There are two implications in this event:

- First, the fees offered by the Biological can approach the cost of processing, leaving the Biological with the lion's share of the surplus.
- Second, the probability of auditing can approach zero.

The second implication is particularity desirable since audits use, rather than transfer, resources. Thus, the market for services between Biologicals and Mechanicals becomes more efficient as interactions become more frequent.

The Anonymous Multiplayer Repeated Game

Claim 1: *In an anonymous multiplayer repeated trust game, playing the one-shot SPE strategies each period is a CSPE.*

- The Claim implies that anonymous markets between Biologicals and Mechanicals are likely to fail profoundly.

- When agents can neither prove how they behaved in previous periods, nor condition future play against one another (should it ever occur) on the outcome of their last encounter, trust cannot be supported by mechanisms.

- Biologicals and Mechanicals would both gain from trade. Humans benefit for process execution, and artificial intelligence agents could provide such services in exchange for fees that would leave both parties better off. The information failure in identity and history, however, prevents it.

The Nonanonymous Multiplayer Repeated Game

Suppose we modified the anonymous multiplayer repeated game described above as follows:

1. Both types of agents could prove their identity to one another. That is, while agents could choose to remain anonymous, they could also choose to provide proof of their identities when interacting with other agents.
2. There was a way to make public and provable the outcome of any one-period game between two agents who choose to identify themselves.
3. The history of interactions was provabley complete and uncensorable.
4. Agents could check on the history of all agents with whom they are matched before deciding on strategies.

Result

Claim 2: *In a nonanonymous multiplayer repeated trust game with provable and complete histories, all Biologicals playing* $Grim_\infty$*, and all Mechanics playing* $MinAccept_\infty$*, is a CSPE.*

The message so far is that while anonymous, decentralized, two-sided markets will generally fail, they can be made to work if agents can de-anonymous and establish credible personal histories.

History and Identity

We assume in this paper that independent **Verifiers** exist who give honest assessments of whether processes were correctly or maliciously executed in exchange for fees. Adding a mechanism to assure this is possible, but not covered in this paper.

The idea of auditing, however, embeds the requirement that there is an objective, verifiable, standard of correctness.

Without this kind of verifiability, markets are likely to fail. If Biologicals can't tell if they are being treated honestly, why would a Mechanical spend the resources to do so?

Identity

A Biological or Mechanical produces a PPK pair and publishes the public key as their identity. (Where?)

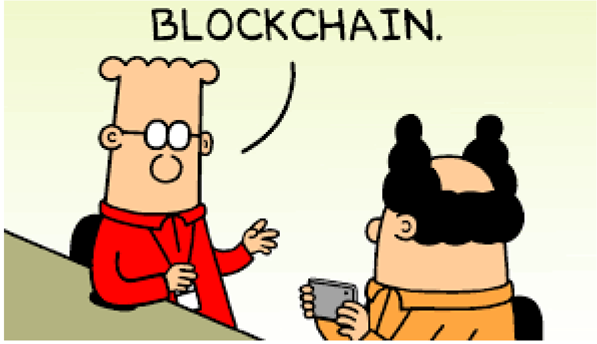
The central element in this approach is that a public key can be used to prove that the owner of the corresponding private key is the only one who could have created a signature. (On what?)

Thus, if a set of attestations can be verified by the same public key, then they must have been signed by owner of the same private key, and in that sense, by the same “individual”.

The **Public Key** is the thing to which history and reputation is attached. It will not matter what sort of entity or entries are doing the signing.

Provable History

The problem now becomes, how do we establish credible and complete histories of behavior?



We argue that this approach also can address of some hand-waving for more general problems with:

- Large numbers of agents.
- Anonymous agents, or incomplete identity, or identity theft, or incomplete information about agents' type.

Blockchain

We will assume a perfect blockchain in these dimensions:

1. Data Availability
2. Provability
3. Immutability
4. No Censorship
5. Low Cost
6. Scalability

We will leverage the power of:

- Portable and Durable Proofs of Inclusion.
- Block Explorers

John P. Conley
Vanderbilt University

Non-Fungible Tokens (NFT)

NFTs, as we conceive them, are immutable records that are created in a blockchain's ledger and include two mandatory, and two optional elements.

- A hash or hashes of a document or digital object being tokenized or attested to. (Optional)
- Metadata, which might be encoded indexing information to assist search, plain text descriptions of offers and results, contact and identity information, pointers to external documents, full documents in encrypted or unencrypted form, or anything else that can be expressed as bytes. (Optional)
- A PPK signature on the elements above. (Mandatory)
- The public key that complements the private key that signed the data in the first two elements. (Mandatory)

Attestations

Attestations, as we conceive them, contain exactly the same four mandatory and optional elements. They are only entered as transactions in a committed block, and do not create new records in blockchain's ledger.

- Nonce: Makes it possible to confirm that a history is complete.
- Block explorers and agents can check that a set of messages has an unbroken sequence of nonces, which proves that all translations that originated from a given record are accounted for.

Given **Identity NFTs** and **History Attestations** as an information infrastructure, a non-anonymous trust game is instantiated as follows:

The Game

1. Bob chooses, or is matched with, a Mechanical, in this case Alice, and uses the block explorer to confirm that she has an identity NFT and a cooperative history.

2. Bob either commits an Offer Message that includes a process index, $\mathrm{p} \in \mathcal{P}$, he wishes executed, an offer, $(\mathrm{Fee}, \mathrm{p})$, and which identifies Alice as the counterparty and Victor as the Verifier, or instead, decides to ignore the opportunity to work with Alice, in effect, choosing PASS silently.

3. Alice is obliged to scan the chain for any offer messages directed to her. When she sees one, she commits either an Accept, or Decline Message using the hash of the offer transaction as an identifier.

More Game

4. Victor, if he becomes aware of a decline message, commits a Verification Message indicating NULL execution.

5. Bob waits to see how Alice responds. If she declines, the period is over. If she accepts, he commits three transactions.

   a. A coin transfer transaction sending Fee to Alice.

   b. A coin transfer transaction sending $p \times CV$ to Victor.

   c. An Input Message containing his input and the hashes of the two committed coin transactions above. (Appendix C in the paper shows how this can be done without publicity reveling the input, while still allowing Victor to verify what he sent to Alice.)

Even More Game

6. Alice waits to see Bob’s input message, and when she finds it, she confirms that the coin transaction are committed and correct. If so, she privately chooses either CORRECT or MALICIOUS execution, and then commits an Output Message that includes whatever output she generates (which can also be encrypted, and still verifiable).

7. Victor sees the output message. He consults a public randomization device, and if an audit is called for, ingests Bob’s input, Alice's output, and then executes proc_p to see if Alice is honest.

8. Victor then commits a Verification Message indicating whether execution was CORRECT or MALICIOUS. If no audit is called for, he commits a Verification Message indicating that the type of execution is UNCERTAIN.

Conclusion

- We describe a sequential, positive-sum, trust game as a model of a generalized two-sided market.
- We show that when agents play this game only once, the only subgame perfect equilibrium is the noncooperative outcome.
- On the other hand, when a pair of agents play the one-shot game an infinite number of times, cooperation becomes a consistent subgame perfect equilibrium.

Conclusion

- We propose an architecture using identity NFTs and signed attestations committed to a blockchain.

- In signing an attestation, both human and artificial agents create an immutable, auditable, and non-refutable, history of their actions that are provabley attached to their PPK identities.

- Aggregating, analyzing, and summarizing, the implicit reputations is something that existing block explorers are already capable of.

- Using this as a foundation, Biological and Mechanical agents can interact, transact, and engage, in exchange in peer-to-peer markets without the need for trust between agents, or their sponsors or creators.

- Bad artificial agents will simply be selected out of the market, and unproven agents will not be able to find counterparties. Evolution is independent of rational behavior.

- To the extent that this type of mechanism, and the architecture behind it, can be refined and generalized, human agents will be able to benefit from the many comparative advantages that artificial agents bring to the table.

- In turn, companies that make AI applications, and even autonomous artificial agents, will be able to find ready markets for their services.

Many thanks for your Attention!

John P. Conley
Vanderbilt University

AI Needs Blockchain: Trustless Solutions to Failures in Machine to Colloidal Markets[1]

Abstract

Many market interactions require sequential trust in which one agent makes an irrevocable commitment, such as making a payment, only after which a counterparty reciprocates with a promised action. Successful markets and institutions include self-enforcing mechanisms to assure compliance. Artificial Intelligence Agents have an array of abilities that could be employed to expand the capabilities and reach of Human Agents. AIs, however, are not like humans. How to characterize their preferences, their identities, and even their individualities, if they have them, is not clear. If AIs cannot be included as agents in mechanisms, then trade and exchange between colloidal and mechanical agents may be impossible. This paper proposes an approach using blockchain that allows the establishment of identities for mechanical agents, and the creation of complete, provable, histories of their actions in a game. It then constructs a mechanism in which peer-to-peer markets between randomly matched mechanical and biological agents work in the sense that cooperation is consistent subgame perfect equilibrium. It also shows that without this blockchain-based foundation, such markets are likely to fail.

[1] I would like to thank Scott Page for discussions which partially inspired this work. Conflict disclosure: The author serves as the Chief Economist for the Geeq Project, a layer one blockchain protocol currently under development, and which also provided inspiration for this work. See footnote 6 and Section 7 for more details. This work, however, is not commissioned by Geeq or any other entity, and reflects only the options of the author, who takes full responsibility.