AI needs blockchain:[1] Trustless solutions to failures in machine to colloidal markets

John P. Conley[2]

Vanderbilt University

February 2024

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 individuality, 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.

Keywords: Artificial Intelligence, Blockchain, P2P Markets Two-sided Markets, Machine to Colloidal Markets, Mechanism Design, Identity, Public Key Encryption

(Notes: A truncated version omitting formal treatment of the game theory included here is published as John P. Conley (2024) "AI Needs Blockchain: Trustless Solutions to Failures in Machine to Colloidal Markets" *MARBLE Conference Proceedings*, (5th International Conference on Mathematical Research for Blockchain Economy), Forthcoming

[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 8 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.

[2] Department of Economics, Vanderbilt University, Nashville, TN 37235, j.p.conley@vanderbilt.edu.

1. Introduction

Artificial Intelligence is here. What this means for human society in unclear. Machines either can, or very shortly will, pass the Turing test. Whether they have, or ever will, develop true sentience is an open question.

Whatever the case, AI is certainly more efficient at accomplishing many types of tasks than humans, and this set will expand rapidly in the coming few years. The rate at which AI displaces human labor in entire categories of work may cause dislocation on a scale never seen before (Acemoglu and Restrepo 2018; Trammell and Korinek. 2023; Zarifhonarvar 2023).

On the brighter side, AI's can also assist humans by taking over some of the more tedious aspects of work, and allow humans to focus instead on those that require judgment, creativity, intuition, and especially, trust (Babina, Tania, et al., 2024). As a technology, AI can magnify human potential, and extend it in directions we have yet to even contemplate.

There is a large and growing literature on machines as participants in games related to financial markets, (Bebeshko, et al. 2022) oligopoly pricing (Calvano et al. 2020), auctions (Bichler et al. 2021) learning (Zeng, et al. 2021), many other things. It is tempting to anthropomorphize artificial intelligences as just another economic actor, although with a very different cognitive profile than humans. The question then becomes, can mechanical and biological agents find a way to cooperate and work together? How can the gains from trade in machine to colloidal markets be realized, and what problems are we likely to encounter?

Mechanism design, particularly market design, is the natural place to look for answers. The literature is limited, and does not seem to address such questions as two-sided or peer-to-peer markets between humans. While there are clearly gains to trade, it becomes immediately clear that AIs cannot simply be slotted in as ordinary actors. See Sima, Violeta, et al. (2021) for an extensive discussion of human-machine interaction. We argue that there are three central reasons for this.

First, economic actors are individuals. Even if agents in a game are anonymous with respect to one another, they all retain a sense of their own individuality. Individuality might be thought of a continuity of consciousness, which in humans, creates a continuity of preferences, memory, and concern about an individual's future. Humans certainly change over time as preferences evolve and memories fade. Such changes, however, take place in ways, and at a pace, that human societies understand and incorporate into their institutions.

It seems unlikely that a non-sentient machine intelligence would be able to conceive of itself as an individual. It is unclear if even sentient machines would do so. AI's can be created at will, copied and cloned without loss, and altered in fundamental ways by changing algorithmic parameters or the data that the machine has available. Are AI twins the same individual? Are they different if some parameter changes? If so, what level of change in an instance of an AI is sufficient to break the continuity of conciseness, assuming it exists at all?

Economists model individual humans as agents who have preferences and constraints. Can machine intelligences, even sentient ones, have preferences? What does an AI want? (See Gabriel 2020 for some speculations.) Perhaps its preferences are identical to the colloidal who created it. This seems unlikely simply because such preferences would have to be encoded on a physical platform with very different processes, cognitive speed, memories, and so on. A creator might try to teach, but what the student learns is only an echo. Modeling AIs as decision theoretic also seems problematic since they learn, grow, and change, in unpredictable ways over time.

Second, at least in real (as opposed to virtual) space, human agents have evolved many ways of identifying and differentiating individuals. It is possible to fool us, but changing appearance, knowing enough about a person's history, and learning how to act as they would act, is a difficult task for an impostor with human limitations.

In virtual space, proving identity becomes much more difficult. We rely on the trinity of something you know, have, and are, in various combinations. Unfortunately, people forget what they know, and bad actors learn and remember. Phones, ID cards, and similar objects, can be copied or stolen. Biometric approaches can be spoofed, are invasive, and are often difficult to use. AI's employed as bad actors will make all these methods less secure in the future.

Machine intelligences can share and clone knowledge, and since anything they have is virtual, "objects" can be shared and cloned as well, and ultimately, we don't know what they really are in the first place. It seems we would have to address the question of what exactly an individual is before we can assign, much less prove, an identity for a machine intelligence.

Third, humans decide who to trust on the basis of the reputation. In turn, reputation depends on the history of actions of agents. Credible sources of information are essential. How we extend the idea of trust to machines given their differences? See Glikson and Woolley 2020, Oksanen, et al. (2020) and Lockey, et al.(2021) for recent discussions of empirical and experimental work regarding human trust in machines.

It is probably more accurate to say that humans don't really trust at all. Instead, we rely on social mechanisms to enforce good behavior. Behaving honestly[3] over a long period of time requires forgoing many opportunities for short-term gains. Societies penalize those who are caught being dishonest. In some cases, this is a collective punishment. In many cases, however, punishment takes the form of independent rational decisions on the part of individual members of the society to refrain from interacting with, or "trusting", agents who have a history of dishonesty.

Social mechanisms like these depend on having reliable information. First-hand observations are best, but second-hand reports from "trusted" agents are also useful. Confidence that the information is relatively complete is also important. People are rightly suspicion of gaps in CVs or employment history.

[3] We use "honesty" as a shorthand for conforming to social expectations in interaction, and thereby avoiding censure.

Critically, such mechanisms rely on a high likelihood of future interactions. If a dishonest agent can simply leave town and start over, sanctions are meaningless. This is probably the main reason that we are more likely to trust people in our own family, tribe, profession, and social, ethnic, or religious group. The inside options for interactions with members of one's own group are more attractive than the outside options, given such trust structures.

Without identity, there is nothing to which a history can be attached. Without history, there is no reputation to evaluate. Without individuality and continuity, it is not clear if the notion of repeated interaction is even meaningful. Without any of these, how can one design mechanisms that create the kind of "trust" required to support machine to colloidal markets? And without such markets, how can we realize the enormous gains from trade that interactions with AIs promise?

The paper proceeds as follows. Section 2 defines a sequential trust game in which a biological is the first mover who decides whether to give a mechanical a fee in exchange for assistance. Having received the fee, the mechanical decides to execute the process either correctly, or maliciously. The biological cannot force the mechanical to behave honestly, and so must hand over its fee "trusting" in the promise of good behavior by the mechanical. As he hands over his fee, however, the biological commits to a probability of running a costly audit to determine the correctness of the output received, and so the honesty of the mechanical.

Section 3 show that if the trust game is played only once, the market fails in the sense that the mechanical is always malicious if the biological makes an offer, and so the biological chooses to pass on the opportunity instead. Mutually beneficial cooperation between the biological and mechanical is impossible in this case.

Section 4 considers an infinitely repeated trust game played each period between one biological and one mechanical. We show that when the agents play grim trigger-like strategies, cooperation becomes possible. In addition, the first-mover advantage allows the biological to force the outcome into an equilibrium that minimizes both the fees paid, and the probability of the required audit.

Section 5 shows that the result in Section 4 falls apart when there are many agents on each side of the market who are randomly, and anonymously, matched. Since agents cannot provably identify themselves to one another, they are also unable to keep meaningful histories of previous interactions. As a result, the environment devolves into a series of unconnected one-shot games, and only the noncooperative outcome is possible.

Section 6 shows that the result in Section 4 is recovered if agents have a method of proving their identity to one another, and if a complete and provable history of the outcomes of all interactions between randomly matched agents is known to all. Provable identity and history fixes the market failure found in the anonymous agent case, and it becomes possible for human and artificial agents to transact, interact, exchange, and create value, without the need for a trusted intermediary. The key is that, just as in traditional human interactions, trust is not needed. Identity and history allow the creation of mechanisms that make good behavior incentive compatible, or more precisely, a consistent subgame perfect equilibrium.

Section 7 develops an architecture using public/private key cryptograph and blockchain that provides the required foundation for mechanisms described in Section 6. This architecture uses NFTs as to create PPK identities, and signed attestation transactions for communications that create provable histories. We show how this approach obviates the need to engage the question of individuality for machine intelligence, sentient or otherwise. Identity is private key, and the nature of the agent who knows it is unimportant. The preferences of mechanicals, how they might be formed, and even their existence, is also unimportant. What matters is behavior. Mechanicals that don't behave honestly are ignored by biologicals, and in a sense, selected against in an evolutionarily dynamic. Section 8 concludes.

2. The Model

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

$$\begin{array}{rl} \text{Biologicals:} & b \in \{1,\ldots B\} \equiv \mathcal{B} \\ \text{Mechanicals:} & m \in \{1,\ldots m\} \equiv \mathcal{M}. \end{array}$$

Mechanicals have a comparative advantage at executing certain types of tasks, booking airline reservations, filing taxes, or optimizing investment portfolios, for example. We call each of these tasks a **Process**, which from a formal standpoint is a mapping from inputs to outputs:

$$\textbf{Proc}: \text{INPUT} \Rightarrow \text{OUTPUT}$$

where

$$\text{Proc}_p \in \{\text{Proc}_1,\ \hat{u}\ \ \text{Proc}_P\} \equiv \text{PROC}$$
$$\text{input}_i \in \{\text{input}_1,\ \hat{u}\ \ \text{input}_I\} \equiv \text{INPUT}$$
$$\text{output}_o \in \{\text{output}_1,\ \hat{u}\ \ \text{output}_O\} \equiv \text{OUTPUT}$$

and

$$p \in \{1,\ \hat{u}\ \ P\} \equiv \mathcal{P},\ i \in \{1,\ \hat{u}\ \ I\} \equiv \mathcal{I},\ \in \{1,\ \hat{u}\ \ O\} \equiv \mathcal{O}.$$

■

Just as executing processes is difficult for a Biological, verifying that a Mechanical has executed a process correctly is also costly. A **Verification** is a mapping from processes, inputs, and outputs, to a truth value.

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

such that

$$\forall\, p \in \mathcal{P},\ i \in \mathcal{I},\ \text{and } o \in \mathcal{O}$$
$$\text{Verify}(\text{Proc}_p, \text{input}_i, \text{output}_o) = \text{CORRECT}\ {}_{c}\quad \text{Proc}_p(\text{input}_i) = \text{output}_o$$
$$\text{Verify}(\text{Proc}_p, \text{input}_i, \text{output}_o) = \text{MALICIOUS}\ {}_{c}\quad \text{Proc}_p(\text{input}_i) \neq \text{output}_o$$

■

Audits are conducted by external agents called **Verifiers**, which are not explicitly modeled in the current paper, and who are assumed to be honest. Verifiers are paid in advance for a probabilistic audit that depends on a public randomization device.

For example, if an audit costs \$10, a Biological would send a Verifier \$1 in exchange for an audit executed with a 10% probability. If the public randomization device indicates that an audit is required, the Biological sends the Verifier the inputs he provided to the Mechanical, and the outputs he received in return. The Verifier then runs the relevant process itself, and reports whether the Mechanical chooses CORRECT or MALICIOUS execution. We discuss the meaning of audit, verification, and provability, in more detail in Section 7.

Let $\mathrm{CP} \in (0, \overline{\mathrm{CP}}]$ denote the **Cost of Executing a Process** correctly to a Mechanical:

$$\textbf{CostProc}: \mathrm{PROC} \Rightarrow (0, \overline{\mathrm{CP}}].$$

Let $\mathrm{CV} \in (0, \overline{\mathrm{CV}}]$ denote the **Cost of Verifying an Execution of a Process** to a Verifier:

$$\textbf{CostVerify}: \mathrm{PROC} \Rightarrow (0, \overline{\mathrm{CV}}]$$

Biologicals and Mechanicals play a sequential **Trust Game** in which 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:

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

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

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

which is a binding commitment if the offer is accepted. If a Biological decides to PASS, he does not send the Mechanical any inputs.

The 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 $(\mathrm{p} \times \mathrm{CV})$ 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 period 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.

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

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

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

■

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. Biologicals do know that correctly executed processes increase their welfare, but are unable to separate this contribution from the many other, difficult to understand, events that affect them positively and negatively.

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

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

where if

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

then

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

■

While Mechanicals do not have utility functions in the same sense as Biologicals, we will assume that they maximize a payoff function that depends on fees collected, and how processes were executed. 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.

MALICIOUS execution gives Mechanicals a higher payoff, all else equal, for two reasons. First, it may be that taking the Biological's inputs and executing a different process directly benefits the Mechanical (investing the Biological's funds in assets that pay commissions, for example). Second process execution is costly, and so not executing any process and returning an invented output always gives a higher payoff than CORRECT execution.

Let $\mathrm{MV} \in (0, \overline{\mathrm{MV}}]$ the **Net Value of Malicious Execution** to a Mechanical:

$$\textbf{MaliciousValue}: \mathrm{INPUT} \Rightarrow (0, \overline{\mathrm{MV}}]$$

We interpret this as the net value, including the cost of executing any process it chooses, to the Mechanical. We bound malicious value away from zero to reflect the idea that a Mechanical can always choose not to execute any process after accepting an offer, but still gets at least some value from seeing the Biological's input.

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

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

where

(1) if

$$\mathrm{a}^b = (\mathrm{Fee}, \mathrm{p}) \text{ and } \mathrm{a}^m = \mathrm{CORRECT}$$

then

$$F^b(a^b, a^m) =$$

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

$$F^m(a^b, a^m) = \text{Fee} - \text{CostProc}(\text{Proc}_p)$$

and

(2) if

$$a^b = (\text{Fee}, p) \text{ and } a^m = \text{MALICIOUS}$$

then

$$F^b(a^b, a^m) = -\text{Fee} - p \times \text{CostVerify}(\text{Proc}_p) - \varepsilon$$

$$F^m(a^b, a^m) = \text{Fee} + \text{MaliciousValue}(\text{input}_i)$$

and

(3) if

$$a^b = (\text{Fee}, p) \text{ and } a^m = \text{NULL}$$

then

$$F^b(a^b, a^m) = 0$$

$$F^m(a^b, a^m) = 0$$

and

(4) if

$$a^b = \text{PASS} \text{ and so } a^m = \text{NULL}$$

then

$$F^b(a^b, a^m) = 0$$

$$F^m(a^b, a^m) = 0$$

■

Note that we subtract ε from the payoff to a Biological when it makes an offer which is accepted, but where the Mechanical chooses MALICIOUS execution. This reflects the small cost of transmitting the input to the Mechanical.

Since fees and audit probabilities are not bounded away from zero, this cost serves to make Biologicals prefer to PASS rather than send a trivial offer, $(\text{Fee}, p) = (0, 0)$, to Mechanicals if they know it will result in MALICIOUS execution. This ε cost is incorporated directly into the Biological's utility function in the event of CORRECT execution.

3. The Two-Player One-Shot Game

We first consider the case where one Biological one Mechanical play the sequential trust game described above once.

A **Strategy for a Biological** is a choice from his action space: either to offer the Mechanical a non-negative fee with some probability of verification to execute a process, or to PASS:

$$s^b \equiv \mathcal{A}^b \equiv \mathcal{S}^b$$

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.

$$s^m : \mathcal{A}^b \Rightarrow \mathcal{A}^m$$

and

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

■

A **Strategy Profile** is denoted:

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

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

Given some $(\text{Proc}_p, \text{input}_i) \in \text{PROC} \times \text{INPUT}$, 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.

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:

$$s^b = \text{PASS} \text{ and } s^m(\text{Fee}, p) = \text{MALICIOUS},$$

and

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

Proof:

This, and all other proofs, are contained in the Mathematical Appendix.

■

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.

4. The Two-Player Repeated Game

Next we consider the case where one Biological one Mechanical play the sequential game an infinite number of times in succession. Thus, each agent, $x \in \{b, m\}$, chooses an action in each period, $t \in \mathcal{T}$:

$$a_t^x \in \mathcal{A}^x$$

A **Sequence of Realized Actions** is denoted:

$$(a_0^x,\ \hat{u}\ \ a_t^x) \equiv A_t^x \in \underbrace{\mathcal{A}^x \times \hat{u}\ \times \mathcal{A}^x}_{\text{t times}} \equiv \mathcal{A}_t^x$$

where

$$\mathcal{A}_t^x \subset \mathcal{A}_\infty^x \equiv \mathcal{A}^x \times \mathcal{A}^x \times \ldots$$

and

$$x \in \{b,\ m\}.$$

■

Note that A_t^x has $t + 1$ elements.

In each period, one of four observable **Events** occurs:[4]

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.

The actions chosen by agents in period t result in an event being realized at the end of each period. The sequence of events from period 0 to period t, therefore, define the game's history as of the beginning of the next period, $t + 1$. Formally,

$$\forall\ t \in \mathcal{T}$$

$$h_t \in \mathcal{H} \equiv \{\text{COR},\ \text{MAL},\ \text{UNC},\ \text{NUL}\}$$

[4] We could enrich the event space to distinguish the strategy choice pairs (PASS, NULL), and ((Fee, p), NULL), where the Biological passes and so the Mechanical must choose NULL execution. and the Biological makes an offer which is declined by the Mechanical, respectively. We do not do so in the current paper and instead class both events as indicating a noncooperative history. This is because it will not matter for the equilibria we explore here, and so only serves to add complexity. In Section 7, an architecture of messages and actions using blockchain transactions is developed which opens some additional event possibilities such as the Biologicals behaving dishonestly in the sense of making false claims against honest Mechanicals. We may explore these details in future work.

is the event that is realized at the end of period $t - 1$, and so the **Period t History of Play** is:

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

where

$$\mathcal{H}_t \subset \mathcal{H}_\infty \equiv \mathcal{H} \times \mathcal{H} \times \ldots$$

and

$$h_0 \equiv \text{UNC}.$$

■

By convention, the history the beginning of period 0 is defined to be $H_0 = (h_0) = (\text{UNC})$, since there is no period $t = -1$, and therefore no actual event could have been realized. We will, however, consider subgames where H_0 has other values.

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**, Formally,

$$H_t \in \mathcal{H}_\infty^{\text{coop}} \subset \mathcal{H}_\infty$$

where

$$\forall\ t \in \mathcal{T}$$

$$h_t \in \{\text{COR}, \text{UNC}\}.$$

■

In the interest of simplicity, we assume the following for the remainder of the Section:

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

$$\text{CostProc}(\text{Proc}_p) = \text{CP}$$

$$\text{CostVerify}(\text{Proc}_p) = \text{CV}$$

$$\text{MaliciousValue}(\text{input}_i) = \text{MV}$$

and if

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

then

$$\text{Utility}^b(\text{Proc}_p, \text{input}_i, \text{output}_o) = \text{U}$$

In words, we assume that cost of executing and verifying processes, the utility Biologicals receive from CORRECT execution, and the value of MALICIOUS execution to the Mechanical, are all constant in the sense that they are independent of the process, input, and output.

The **Probability Distribution over Events** as a function of actions is defined as follows:

$$\textbf{ProbEvent}: \mathcal{A}^b \times \mathcal{A}^m \Rightarrow \Delta^3 \equiv$$

$$(p^{COR}, p^{MAL}, p^{UNC}, p^{NUL}) =$$

(1) if

$$a^b = (Fee, p) \in [0, \bar{F}] \times [0, 1] \text{ and } a^m = CORRECT$$

then

$$(p, 0, (1 - p), 0)$$

and

(2) if

$$a^b = (Fee, p) \in [0, \bar{F}] \times [0, 1] \text{ and } a^m = MALICIOUS$$

then

$$(0, p, (1 - p), 0)$$

and

(3) if

$$a^b = PASS, \text{ or } a^m = NULL$$

then

$$(0, 0, 0, 1).$$

■

Strategies for the repeated game depend upon history. A **Period t Strategy for Biologicals** is any mapping from period t histories into the Biological action space.

$$\forall\ t \in \mathcal{T}$$
$$s_b^t : \mathcal{H}_t \Rightarrow \mathcal{A}^b$$

and

$$s_b^t \in \mathcal{S}_b^t.$$

■

Biologicals choose an action before Mechanicals. Thus, 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:

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

and

$$\forall\ s_t^m \in \mathcal{S}_t^m,\ \forall\ H_t \in \mathcal{H}_t$$

$$s_t^m(H_t, PASS) = NULL.$$

■

A **Strategy Profile** for the repeated game is denoted:

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

where

$$S_\infty^x \in \prod_{t=0}^{\infty} \mathcal{S}_t^x \equiv \mathcal{S}x_\infty.$$

■

Thus:

- $s^x_t \in \mathcal{S}^x_t$ is a strategy for agent x for period t only.
- $\mathcal{S}^x_t$ is the space of all possible strategies for agent x for period t only.
- $S^x_t \subset \mathcal{S}^x_\infty$ is a list of strategies for agent x for each period from 0 to t, (which is t + 1 periods in all).
- $\mathcal{S}^x_\infty$ is the space of all possible lists of strategies for agent x for all periods (an infinite number).

Note that the Biological only knows 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.

The **Period t Beliefs** are denoted as follows:

$$\forall\ t \in \mathcal{T}$$

$$\beta^m_t \in \mathcal{S}^m_t \text{ and } \beta^b_t \in \mathcal{S}^b_t.$$

■

Note that β^b_t is the Mechanical's belief about the Biological, and the opposite for β^m_t.

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. Thus, we add a consistency condition on 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

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

(1) 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

(2) if

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

then

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

■

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.

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, $(H_T, a_T^b) \in \mathcal{H}_T \times \mathcal{A}^b$.

The expected payoff of strategy choices in any subgame depends on the **Probability of Future Histories**. The conditional probability that $(h_{T+1}, \ldots h_{\bar{T}}) \subset H_{\bar{T}} \subset H_\infty$ will be the future history of events starting from a subgame defined by $H_T \subseteq H_{\bar{T}}$ when agents follow strategies $S_{\bar{T}}^b \in \mathcal{S}_\infty^b$ and $S_{\bar{T}}^m \in \mathcal{S}_\infty^m$ over the interval $t \in (T, \ldots \bar{T} - 1)$ is given by the mapping:

$$\textbf{ProbHist}: \mathcal{T} \times \mathcal{T} \times \mathcal{H}_\infty \times \mathcal{S}_\infty^b \times \mathcal{S}_\infty^m \Rightarrow [0, 1] \equiv \prod_{t=T}^{\bar{T}} p_t^{\bar{h}}$$

such that

$$p_T^{\bar{h}} = 1 \text{ if } \bar{h} = h_T \in H_T$$
$$p_T^{\bar{h}} = 0 \text{ if } \bar{h} \neq h_T \in H_T$$

and

$$\forall\ t \in (T+1, \ldots \bar{T})$$
$$p_t^{\bar{h}} = p^{\bar{h}} \in (p^{COR}, p^{MAL}, p^{UNC}, p^{NUL}) = \text{ProbEvent}(a_{(t-1)}^b, a_{(t-1)}^m)$$

where

$$a_{(t-1)}^b = s_{(t-1)}^b(H_{(t-1)}) \text{ and } a_{(t-1)}^m = s_{-1}^m(H_{(t-1)}, s_{(t-1)}^b(H_{t-1})).$$

■

Note the following:

- It may not be possible for subgame history H_T to be realized given $S_{(T-1)}^b \in \mathcal{S}_{(T-1)}^b$ and $S_{(T-1)}^m \in \mathcal{S}_{(T-1)}^m$ for periods $t \in (0, \ldots T-1)$. In this case, the unconditional probability of the future history would be zero. The ProbHist mapping, however, gives the conditional probability of the future history for subgames regardless of their likelihood.
- In particular, the last element of the history, $h_T \in H_T \subset H_{\bar{T}}$, that defines the subgame is assumed to occur with certainty, since it is what conditions this probability calculation.
- The probability that the supergame, defined by $H_0 = (h_0)$, will end up with history $H_{\bar{T}}$ in period $\bar{T}$ when agent play strategies $S_\infty^b \times S_\infty^m \in \mathcal{S}_\infty^b \times \mathcal{S}_\infty^m$ is:

$$\text{ProbHist}(0,\ \bar{T},\ H_{\bar{T}},\ S^b_\infty,\ S^m_\infty).$$

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

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

Using this, we can calculate 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:

$$\mathbf{EPO^x}: \mathcal{T}\times\mathcal{H}_T\times\mathcal{S}^b_\infty\times\mathcal{S}^m_\infty \equiv$$
$$\prod_{t=0}^{\infty} r^t \sum_{H_{(T+t)}\in\overline{\mathcal{H}}_{(T+t)}} \text{ProbHist}(T,\ T+t,\ H_{(T+t)},\ S^b_\infty,\ S^m_\infty)\times F^x(a^b_{(T+t)},\ a^m_{(T+t)})$$

where

$$\forall\ t \in \mathcal{T}$$
$$a^b_{(T+t)} = s^b_{(T+t)}(H_{(T+t)}) \text{ and } a^m_{(T+t)} = s^m_{T+t}(H_{(T+t)},\ s^b_{(T+t)}(H_{(T+t)}))$$

and

$$\overline{\mathcal{H}}_{(T+t)} \equiv H_T\times\underbrace{\mathcal{H}\times\hat{u}\ \times\mathcal{H}}_{t\ \text{times}}.$$

■

Note the following:

- $EPO^x(0, H_0, S^b_\infty, S^m_\infty)$ is the expected payoff to agent x of the supergame.
- Discounting begins with the subgame period T. That is, $EPO^x(T, H_T, S^b_\infty, S^m_\infty)$ gives the expected value of the subgame where period T is the current period.

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:

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

■

Give all this, we are at last able to define our equilibrium notion. 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:

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

it holds that,

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

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

where

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

such that the Mechanical's beliefs satisfy:

$$\forall\ T \in \mathcal{T},\ \beta^b_T = s^b_T,$$

and the Biological's beliefs satisfy:

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

such that

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

and

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

■

Beliefs in an CSPE are consistent in the following senses:

- Agents assume that their counterparties will choose the same actions in all similar situations.
- In every period T, agents base their beliefs about the future strategies of their counterparties on the last strategy they played. Note that for Mechanicals, this is the current period, while for Biologicals, this is the previous period.
- Since in period $T = 0$, the Biological has not yet seen any Mechanical strategy being played, he is free to form any beliefs that are payoff maximizing for the Mechanical given the strategy chosen by the Biological.

Given these beliefs, CSPE strategies are payoff maximizing in both the supergame, and every subgame defined by H_T, which may or may not be possible given S_∞.

The Mechanical updates his beliefs about the Biological when it sees the period T strategy being played ($\beta^T_b = s^T_b$). This means that the Mechanical bases its response on both the history of play, and the action chosen by the Biological in period T. Note that while the beliefs are in period T must be consistent, the actual strategies the agents play need not satisfy this condition,

Define the **Grim Trigger Strategy** for the Biological as follows:

$$\mathbf{Grim}: \mathcal{H}_\infty \times [0, \bar{F}] \times [0, 1] \Rightarrow \mathcal{A}^b \equiv$$

$$\mathrm{Grim}_\infty(H_\infty) \equiv (\mathrm{Grim}_0(H_0), \ldots \mathrm{Grim}_t(H_t), \ldots)$$

where

$$\forall\ t \in \mathcal{T}$$

(1) if

$$H_t \in \mathcal{H}_\infty^{coop}$$

and

(a) if

$$\exists\ (Fee, p) = \underset{(Fee,\, p) \in [0,\, \bar{F}] \times [0,\, 1]}{\operatorname{argmin}} \quad Fee + p \times CV$$

such that

$$EPO_C(Fee, p) \geq EPO_D(Fee, p)$$
$$U - (Fee + p \times CV) \geq 0$$

then

$$a_t^b = (Fee, p)$$

and

(b) if

$$\nexists\ (Fee, p) = \underset{(Fee,\, p) \in [0,\, \bar{F}] \times [0,\, 1]}{\operatorname{argmin}} \quad Fee + p \times CV$$

such that

$$EPO_C(Fee, p) \geq EPO_D(Fee, p)$$
$$U - (Fee + p \times CV) \geq 0$$

$$a_t^b = PASS$$

and

(2) if

$$H_t \notin \mathcal{H}_\infty^{coop}$$

then

$$a_t^m = PASS$$

■

A grim trigger strategy for the Biological attempts to enforce **Cooperation** in the form of CORRECT execution by the Mechanical each period at the least possible cost. MALICIOUS execution, however, cannot be punished unless it is detected by audit. Thus, if the Biological becomes aware of a **Defection** from cooperative behavior, it punishes the Mechanical by choosing to PASS for all future periods. This results in a zero payoff for both agents. The Biological also punishes NULL execution (that, declining an offer), which is always detected without audits, in the same way.

To make the proofs more readable, we define the following shorthand notation. Denote the one period **Defection Payoff** to the Mechanical as:

$$D \equiv (Fee + MV) \geq 0$$

and the one period **Cooperative Payoff** to the Mechanical as:

$$C \equiv (Fee - CP)$$

Note that since $\text{Fee} \geq 0$ and $\text{MV} > 0$, it must be that $\text{D} > 0$. Given this, if the Biological follows the strategy $\text{Grim}_\infty(\text{H}_\infty)$, then the expected payoff to the Mechanical of cooperating and choosing CORRECT execution each period is:

$$\text{EPO}_\text{C}(\text{Fee}, \text{p}) \equiv \sum_{t=0}^{\infty} r^t \times \text{C} = \frac{\text{C}}{(1 - r)},$$

while the expected payoff of choosing MALICIOUS execution each period until a successful audit detects its defection and triggers the Biological to PASS for all future periods is:

$$\text{EPO}_\text{D}(\text{Fee}, \text{p}) \equiv \sum_{t=0}^{\infty} (1 - p)^t r^t \times \text{D} = \frac{\text{D}}{(1 - r + rp)} > 0.$$

The **Minimal Acceptance Strategy** defines the set of offers and histories that are sufficient to convince the Mechanical to choose CORRECT execution. Specifically, the Mechanical assumes that the Biological will take the same action (an offer or PASS) given similar histories in each period. If the expected payoff of choosing CORRECT verses MALICIOUS execution in the current, and all future, periods under this assumption a least as large, the Mechanical chooses CORRECT execution. If not, the Mechanical chooses MALICIOUS execution. Of course, the Mechanical must choose NULL execution if the Biological chooses PASS. Formally:

$$\textbf{MinAccept}: \mathcal{H}_\infty \times \mathcal{A}^b_\infty \Rightarrow \mathcal{A}^m \equiv$$
$$\text{MinAccept}_\infty(\text{H}_\infty, \text{A}^b_\infty) \equiv (\text{MinAccept}_0(\text{H}_0, a^b_0), \ldots \text{MinAccept}_t(\text{H}_t, a^b_t), \ldots)$$

where

$$\forall\ t \in \mathcal{T}$$

(1) if

$$\text{H}_t \in \mathcal{H}^{coop}_\infty$$
$$a^b_t = (\text{Fee}, p) \in [0, \bar{F}] \times [0, 1]$$
$$\text{EPO}_\text{C}(\text{Fee}, p) \geq \text{EPO}_\text{D}(\text{Fee}, p)$$

then

$$a^m_t = \text{CORRECT}$$

and

(2) if

$$\text{H}_t \in \mathcal{H}^{coop}_\infty$$
$$a^b_t = (\text{Fee}, p) \in [0, F] \times [0, 1]$$
$$\text{EPO}_\text{C}(\text{Fee}, p) < \text{EPO}_\text{D}(\text{Fee}, p)$$

then

$$a^m_t = \text{MALICIOUS}$$

and

(3) if

$$\text{H}_t \notin \mathcal{H}^{coop}_\infty$$

$$a_t^b = (\text{Fee}, p) \in [0, F] \times [0, 1]$$

then

$$a_t^m = \text{MALICIOUS}$$

and

(4) if

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

then

$$a_t^m = \text{NULL}$$

■

Note that the Mechanical never declines an offer (that is, chooses NULL when $a_t^b = (\text{Fee}, p)$). This is because doing so results in a zero payoff, is interpreted as noncooperative behavior, and is always detected. Accepting the offer and choosing MALICIOUS execution, on the other hand, gives a positive payoff, and is only interpreted as noncooperative behavior if an audit happens to be run (which happens with a probability of $p \leq 1$).

Give all this, we are at last able to state the main Theorem of this Section. Theorem 2 says that in the two-player repeated game defined above, the strategy profile:

$$S_\infty = (\text{Grim}_\infty, \text{MinAccept}_\infty),$$

is a Consistent Subgame Perfect Equilibrium.

Theorem 2: *If*

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

then

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

is a Consistent Subgame Perfect Equilibrium.

Proof: See the Mathematical Appendix for a series of Lemmas that collectively prove this result.

■

Perrett and Powers (2021) explore a repeated game between human and artificial agents in an evolutionary context that provides an interesting contrast. They find that agents eventually do not seek full information about the history of play, but end up simply checking periodically. Even periodic monitoring, however, presupposes that a human's counterparty has an identity. We will see in the next Section that unless machine identity has a clear foundation, cooperation seems to be impossible when agents are fully strategic.

4.1. Discussion

There are two possible histories that can evolve in equilibrium depending upon whether the parameters of the game allow a solution the minimization problem that defines the Biological's grim trigger strategy. Formally, upon whether:

$$\exists\ a_t^b = (\text{Fee}, p) = \underset{(\text{Fee},\, p) \in [0, \bar{F}] \times [0, 1]}{\text{argmin}} \quad \text{Fee} + p \times CV$$

such that

$$\text{EPO}_C(\text{Fee}, p) \geq \text{EPO}_D(\text{Fee}, p)$$
$$U - (\text{Fee} + p \times CV) \geq 0.$$

If there does, then the Biological offers the solution, $a_b = (\text{Fee}, p)$, to the Mechanical, and the Mechanical, using the minimum acceptance strategy, responds with CORRECT execution. As a result, the event observed at the end of the period is $h \in \{\text{COR}, \text{UNC}\}$, depending on whether an audit takes place. Whatever the outcome, this leads to cooperative history in all periods, and for the entire future.

If no solution exists, then in period $t = 0$, the Biological chooses PASS, the Mechanical responds with NULL (the only possible choice), and the event observed at the end of the periods is $h = \text{NUL}$. This leads to noncooperative history in period $t = 1$, and for the entire future.

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

$$\text{EPO}_C(\text{Fee}, p) \equiv \frac{\text{Fee} - CP}{(1 - r)} \geq \frac{\text{Fee} + MV}{(1 - r + rp)} \equiv \text{EPO}_D(\text{Fee}, p).$$

We find that satisfies all of our intuitions over fee and audit structure.

- $\text{Fee} \geq CP$. That is, fee must always cover the cost of processing. Otherwise, since $\text{Fee} + MV > 0$, the inequality could not be satisfied.
- $CP \uparrow$, or $MV \uparrow$, implies either $\text{Fee} \uparrow$, 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.
- $p = 1$ 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 $\text{Fee} - CP \rightarrow 0$. 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.

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.

5. The Anonymous Multiplayer Repeated Game

Suppose that there are multiple Biologicals and Mechanicals, each of whom is randomly matched to an anonymous counterparty agent each period, and then plays the one-shot game. Assume that the numbers of each type are equal:

$$\begin{array}{rl} \text{Biologicals:} & b \in \{1, \ldots B\} \equiv \mathcal{B} \\ \text{Mechanicals:} & m \in \{1, \ldots m\} \equiv \mathcal{M} \end{array}$$

where

$$B = M.$$

Since agents are anonymous, the history of play does not describe interactions with any specific individual counterparty agent. Rewards and punishments for good and bad behavior based on history, therefore, cannot be correctly targeted.

If a Biological ever makes an offer to a Mechanical to execute a process, it is almost a dominant strategy[5] for Mechanical to choose MALICIOUS execution. In effect, each period is just like a new one-shot game with a counterparty that has not been provably encountered before. The next Biological that the Mechanical encounters at best will condition his strategy on the behavior of the previous Mechanicals he has encountered, not on the unknown behavior of the current one. Given this, it is a best-response for the Biological to choose PASS each period.

This leads to the following Claim:

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

5.1. Discussion

If this Claim is true, it means that anonymous markets between Biologicals and Mechanicals are likely to fail profoundly. When agents can neither prove to 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.

It is true that Biologicals could collect statistical histories regarding the behavior of the anonymous Mechanicals they happened to have encountered. This might even prove useful if Mechani-

[5] See the discussion below for some unlikely interpretations of the generalized game where this might not be a dominant strategy.

cals were exogenously fixed, decision theoretic, types, such as blockchain's Byzantine or non-Byzantine nodes. Such a world, however, seems unlikely. Even if Mechanicals were non-strategic, contrary to the current model, it would be profitable for bad-actors to spin-up Byzantine Mechanicals to harvest fees from credulous Biologicals.

Alternatively, one could imagine a case in which all Biologicals informed one another of each event they encounter as it happens each period. Such universally informed Biologicals could then use a meta-grim trigger strategy where they made offers until any Biological encountered a defecting Mechanical. It might be possible to support a kind of Cooperative CSPE outcome in this case.

We do not explore or formalize this possibility for three reasons. First, even if such equilibria existed, they would be fragile, especially with large numbers of agents, and would not exist at all if new Mechanicals could enter the game. Second, the information requirements would be large. Third, Biologicals would have the trust in the honesty of all other Biologicals to report outcomes correctly.

What this suggests is that trust deficits between Biologicals and Mechanicals may limit the positive impact, not to mention, the market penetration, of coming AI technologies.

6. 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 provably complete and uncensorable.
4. Agents could check on the history of all agents with whom they are matched before deciding on strategies.

Two-sided markets are often mediated through trusted platforms. For example, see Zhou (2017) and Tan, et al. (2020) among many others. In contrast, we consider decentralized two-sided markets with random or endogenous matching.

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

We will state this as a formal theorem in future versions, but doing so requires reworking the model given in Section 5 in the obvious ways to account for multiple agents. (AI would be much faster at generating this analog.)

In any event, to see why this Claim is true, suppose that Biological followed the same grim trigger strategies with the modification that Biologicals base their strategies on the history of a Mechanical in all of it previous interactions. That is, Biologicals never make offers to Mechanicals that have ever declined an offer, or been caught choosing MALICIOUS execution, in any period, with any Biological.[6]

Note first that in period $t = 0$, the no agent has a history. If the costs and other parameters of the game allow a Biological to make an offer as defined by Grim_∞ he does so. In this case, the offer will satisfy:

$$\mathrm{EPO}_C(\mathrm{Fee}, p) > \mathrm{EPO}_D(\mathrm{Fee}, p)$$

and so the Mechanical, following $\mathrm{MinAccept}_\infty$, accepts and chooses CORRECT execution. The same pattern is repeated in every subsequent period. If the Biological does not make an offer under Grim_∞, then the future history is noncooperative, just as in the two-agent game.

On the other hand, a Biological, encountering a Mechanical who has defected in the past, would not choose to make an offer. Remember that Biologicals take as fixed the strategies of all other agents, including other Biologicals. Since the Biologicals that are matched with this Mechanical in all future periods choose PASS, the value of the continuation game for the defecting Mechanical is zero whatever it chooses in the current period. Thus, the Mechanical will always choose MALICIOUS execution if the current Biological makes an offer. As a result, the current Biological is better-off and following Grim_∞ and choosing PASS.

7. History and Identity

The message of the previous Sections 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.

We will assume for now 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. For example, in the case of blockchains with deterministic protocols, it should be the case that given the current ledger state, a proposed block is either valid or invalid. It may also be that given a set of financial inputs, a tax return is, or is not, correct, or is, or is not, op-

[6] The next Section describes an information structure that supports such strategies without burdens on agents.

timized to a certain standard, or that an investment portfolio was, or was not, managed under some specific accepted standard of best-practice.

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? If bots or malicious humans can leave what amount to fake Yelp reviews and have them taken as history, then dishonest Mechanicals can falsely pump their reputations while smearing honest ones. If truth is not provable, then it may as well not exist from a mechanism design standpoint. For example, see Ball and Kattwinkel (2019) who explore a mechanism with probabilist verification of truthful binaries and the impact on the distribution of surplus in the context of identity and authorization.

In this Section, we will assume that truth is provable using Verifiers and develop an architecture that relies on **Public/Private Key (PPK) Cryptography** for identity, and **Blockchain** for histories. It is important to note that our proposal uses blockchain purely as a data source. This contrasts with the standard approach of building decentralized markets using smart contracts. See AlAshery et al. (2020) for energy markets, Hua, et al. (2020), for carbon markets, and Schär, (2021) for financial markets built on smart contracts, for examples.

7.1. Artificial Identity

The philosophical question of whether an artificial intelligence, or other Mechanical, has an identity, much less an individuality, is a difficult one. AIs are distributed over clusters of computers. New instances can be deployed and taken down at will. Exact copies an AI's code and data can be produced, shipped, and then installed, remotely. AI's also change continuously as they ingest and process new data. Can such an agent, even if identified, be punished, and would it care?

Fortunately, we do not need to engage these weighty questions. Instead, we propose that identity is equivalent to a PPK pair. This is by no means a new idea, and the technology is well-known. In the interest of clarity, let us briefly review.

Public and private key pairs are mathematically entangled, asymmetric encryption keys. For our purposes, their essential feature is that anything encrypted with one key in a pair can only be decrypted with its complementary key. Public key encryption is what enables HTTPS, blockchain, digital signing of documents, and many other building blocks of modern information technology.

As an identity for agents, it works as follows. A Biological or Mechanical produces a PPK pair and publishes the public key as their identity. The complimentary private key is kept secret, and used to cryptographically sign attestations that signify agreement to, or responsibility for, certain actions. This might include receiving specific data, making a request for processing, claiming that input was processed incorrectly, or challenging such a claim.

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 the signature. 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".

7.2. Provable History

As we discuss in the introduction, without identity, there is nothing to attach a history of behavior to. Anonymous agents can't establish reputations, nor can they be held accountable for their actions. With identity, it becomes possible to create intertemporal mechanisms to incentivize good behavior.

The problem now becomes, how do we establish credible and complete histories of behavior? This may seem especially challenging when there are many Biological and Mechanical agents in market, and so matches may happen many times per second. Artificial intelligences might be able to handle this volume of information, but it seems like it would be beyond the capacity of a human. The inputs and outputs may also be very large byte strings, and processing, as we mention, could be complex and costly. Finally, how would the Biological know that it had access to all reports, both of good, and bad, behavior?

The solution we propose relies on blockchain. An immediate question is: what blockchain? There are thousands of implications with different consensus mechanisms, security guarantees, costs, scalability, and so on. Rather that answering this question specifically, we give a list of the requirements a blockchain implementation should satisfy for our purposes.

1. **Data Availability**: All inquiries to block explorers regarding transaction and ledger data in particular must be answered correctly.
2. **Provability**: The data provided by block explorers should allow agents to independently prove the correctness, contents, and inclusion of transactions in committed blocks, as well as the state of the ledger at any block height.
3. **Immutability**: All committed blocks (perhaps after a delay) are considered finalized, and cannot be reorganized or otherwise altered.
4. **No Censorship**: All valid transaction requests sent by Biologicals or Mechanicals must be processed by the network, and included in committed blocks without unreasonable delay.
5. **Low Cost**: The cost of having a transaction included in a block must be low relative to the payoff and cost values of the economic environment described above.
6. **Scalability**: The blockchain must have the capacity to include transactions at the scale required by the economic environment described above.

We will assume a perfect blockchain in these dimensions: all valid transactions are immediately, and immutably, included in the next block at zero cost, and all agents in the game are aware of the contents of all blocks. Exploring the impact of less than perfect or manipulable blockchains is a task for another paper.

7.3. Attestations and NFTs

We require one type of record, and one of transaction, to create identities and histories, although there are probably many alternative approaches that would also serve. These are **Non-Fungible Tokens** (**NFT**) and **Attestations**. We will also make use of ordinary coin transactions.

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, as we conceive them, contain exactly the same four mandatory and optional elements. They are only entered as transactions in a committed block, however (if they satisfy the protocol's definition of correctness[7]), and do not create new records in blockchain's ledger. They also include a **Nonce** that makes it possible to confirm that the 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.

In general, attestation transaction and NFT records are not datagram types that are native to blockchains (Hardjono and Smith 2021; Wang, et al. 2021). Instead, they are instantiated using smart contracts. This is problematic because these datagrams, and proof of their ownership, contents, and origin, are only implicit in the smart contract's state. Verification requires rerunning every transaction that targeted the smart contract since it was deployed in the correct sequence. This makes sufficient data availability burdensome, and provability costly.

Using smart contracts also significantly increases costs and limits scalability. As an unhappy bonus, smart contracts have proven to be a significant attack surface for blockchains. See Chaliasos, et al. (2023) or Zhang, et al. (2022) for example. Fortunately, it is possible to implement attestations and NFTs nativity, visibly, and provably.[8]

[7] Correctness under blockchain protocol requires such things as a correct signature, correct nonces, and sufficient funds to pay for a transaction. It has nothing to do with the correctness or content of an attestation message in the context of the game's messaging rules.

[8] Full disclosure: The author is the Chief Economist of the Geeq Project, a layer one blockchain protocol that in fact does instantiate attestations as transactions signed by coin account owners and places them directly in blocks. Geeq's blockchain incorporates NFT mint accounts as ledger records that can create the type of signed NFT ledger records as described in this Section as well. Geeq's protocol also satisfies, or approximately satisfies, the six requirements outlined in Section 7.2.

7.4. An Architecture for Identity

Identity is implemented through NFTs. Agents of either type simply mint, or have minted, an NFT record with a public key of their choosing signed by the complementary private key, which only they know. It might or might not be useful for the NFT to include Metadata that describes the agent type, who its sponsor is, what services it provides, how to contact it, and so on, but very little is needed for our purposes. An **Identity NFT** simply puts into the ledger the provable fact that some agent knows both parts of a PPK pair.

The existence of the identity NFT record allows other agents to connect any attestations signed with the associated private key to this NFT record as an identity, and thereby allows the creation of an attributable history. Since NFTs can be burned, agents can remove them if they discover that their private keys have been compromised. Once an NFT is removed from the ledger, the agent who created and signed the NFT bears no responsibility for any future attestations signed by the private key. It is the responsibility of the counterparty agents to confirm that an identity NFT exists for any agent they plan to do business with.

7.5. An Architecture for History

History is recorded through attestations. There are, no doubt, many ways to do this, and different approaches may be more suitable for different applications. In this Subsection we give a sketch of simple set of game messaging rules that map on to the multiagent game outlined in Section 6. This relies on two main elements. The first is the identity NFTs described above. The second are various types of **Attestation Transactions** that work as messages when committed to a blockchain. Section 9 describes a set of cryptographic and blockchain primitives that support the architecture used in this Subsection.

Below, we call the AI Mechanical agent Alice, the human Biological agent Bob, and the Verifier agent Victor. Attestation transactions are essentially metadata packages that are signed with an agent's private key and then committed to a block in a blockchain. They do not create or modify ledger records except to deduct fees from, and increment the nonce of, the sending coin account. We will refer to them as **Messages**, below.

Game Messaging Rules: A simple approach to communications using blockchain transactions.

The Pregame: All agents, of all types, generate a PPK pair and then create and commit an identity NFT to the blockchain ledger that includes their public key, and may include other details such as their agent type.

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, $p \in \mathcal{P}$, he wishes executed, and an offer, (Fee, p), and 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.
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. (Section 9 shows how this can be done without publicity reveling the input, while still allowing Victor to verify what he sent to Alice.)
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 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 inputs, Alice's output, and then executes proc_p to see if Alice is honest. 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.

Section 9 describes how Victor also plays a role in making sure that Alice and Bob take each of these steps, and do them correctly. If they don't, he commits a Verification Message indicating which party is dishonest.

Taken together, at the end of the period, an event has been certified by Victor that creates a period t history of COR, MAL, NUL, or UNC.

8. Conclusion

We propose 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.

We then extend the game to include randomly matched anonymous agents. Perhaps unsurprisingly, the positive result breaks down, and once again, only the noncooperative outcome is an equilibrium. If the randomly matched agents are non-anonymous, and each agent can establish a complete and credible history of his actions in previous periods, however, then the cooperative outcome can be recovered as a consistent subgame perfect equilibrium.

Economic mechanisms with human agents are built on a foundation that assumes that each agent has well-defined preferences. Concomitant with this is an assumption that, while agents may be anonymous with respect to one another, each has an identity known at least to themselves. In turn, this rests on an assumption that agents have an individuality, or a sense of continuity between periods, and so care what happens to them as an individual in the future.

Artificial Intelligence, as a field. is advancing at a frightening pace. We do not know, however, whether AIs have preferences as we understand them. If they do, are they programmed, or do they evolve autonomously? How would we identify an AI as a separate agent when they can be cloned or deployed with minor variations in different locations, on radically different hardware and networks? Do AIs, even sentient ones, have a sense of individuality or continuity of self over time? Without the answers to these questions, how can we use our familiar tools to create mechanisms and markets that include AIs as agents?

We argue in this paper that we can build such mechanisms without having to address these questions. Identity can be assigned through public/private keys without the requirement that it be attached to an actual individual. More importantly, once we have an identity, we have something to attach a history to.

We propose an architecture using identity NFTs and signed attestations committed to a blockchain. In signing an attestation (which might include an offer of a fee for work, or a work product completed), both human and artificial agents create an immutable, auditable, and non-refutable, records of their actions over time that are provably attached to their PPK identities. Aggregating, analyzing, and summarizing the implicit histories is something that existing block explorers already do.

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.

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.

9. Cryptographic and Blockchain Primitives Appendix

This Appendix defines various cryptographic primitives and the basic datagrams used by the blockchain to generate the provable histories our mechanism requires. It also provides more details about the games messaging rules.

9.1. Cryptographic Primitives

Generic data of arbitrary size, including inputs, outputs, and elements of blockchain transactions and records, are called **Byte Strings**:

$$\mathbf{BYTE\_STRING} \equiv \{\mathrm{byte\_string} \in \{0,1\}^{\mathrm{n}} | \mathrm{n} \in \mathbb{N}\}.$$

A **Hash Function** maps a **Pre-image**, which is a byte string of any length, into an approximately uniform distribution of (usually 32 byte) byte strings called a **Hash Digest**.

$$\begin{gathered}\mathbf{Hash}: \mathrm{BYTE\_STRING} \Rightarrow \{0,1\}^{32}: \\ \mathrm{Hash}(\mathrm{pre\_image}) = \mathrm{hash\_digest}.\end{gathered}$$

There are three sets of agents:

$$\begin{aligned}\text{Biologicals}: &\quad \mathrm{b} \in \{1,\ldots\mathrm{B}\} \equiv \mathcal{B} \\ \text{Mechanicals}: &\quad \mathrm{m} \in \{1,\ldots\mathrm{m}\} \equiv \mathcal{M} \\ \text{Verifiers}: &\quad \mathrm{v} \in \{1,\ldots\mathrm{V}\} \equiv \mathcal{V}.\end{aligned}$$

Each agent, of each type, creates a **Public/Private Key Pair**:

$$(\mathbf{pub\_key^x}, \mathbf{pri\_key^x})$$

where

$$\begin{gathered}(\mathrm{pub\_key^b}, \mathrm{pri\_key^b}) \\ (\mathrm{pub\_key^m}, \mathrm{pri\_key^m}) \\ (\mathrm{pub\_key^v}, \mathrm{pri\_key^v})\end{gathered}$$

are PPK pairs for generic Biologicals, Mechanicals, and Verifiers, respectively. As we mention above, anything encrypted with one of the paired keys can only be decrypted with the complementary key. Asymmetric encryption is limited in that the bytes string being encrypted must be smaller than the key size, and the process is relatively computationally intensive.

We will also use **Symmetric Encryption Keys**:

$$\mathbf{sym\_key}$$

that have the property that byte strings of any length can be encrypted and decrypted with the same key at relatively low computational cost.

An **Encryption Algorithm** (systematic or asymmetric) maps **Plaintext** byte strings into **Ciphertext** byte strings using a key:

$$\mathbf{Encrypt}(\mathrm{key}, \mathrm{plaintext}) = \mathrm{ciphertext}.$$

A Decryption Algorithm maps ciphertext byte strings into plaintext byte strings using a key:

$$\mathbf{Decrypt}(\mathrm{key}, \mathrm{ciphertext}) = \mathrm{plaintext}.$$

A **Signature Algorithm** maps a private key and a byte string into a byte string called a **Signature**. In general, the byte string being signed is the hash digest of a byte string of arbitrary length.

$$\mathbf{Signature}(\mathrm{pri\_key}, \mathrm{byte\_string}) = \mathrm{signature}.$$

Finally, a **Signature Check Algorithm** maps a public key and a signature into a truth value:

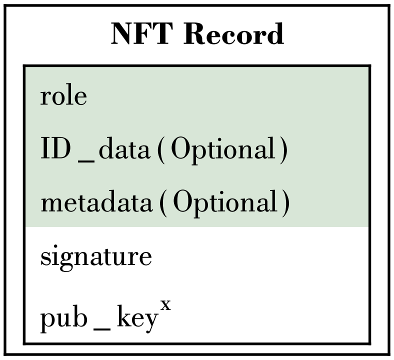
$$\mathbf{SigCheck}(\mathrm{pub\_key}, \mathrm{signature}) \Rightarrow \{\mathrm{TRUE}, \mathrm{FALSE}\},$$

and takes a value of TRUE if and only an agent who had access to $\mathrm{pri\_key}$ created signature, using $\mathrm{byte\_string}$ as the argument.

Given the cryptographic primitives, we construct the following blockchain records and transactions.

9.2. Identity NFTs

Identity NFTs are created by **Mint Accounts**, and are signed by their creator. The three data items (in green) are helpful in the sense that a human looking at such a record would know that a certain public key is associated with a specific agent (Alice, Bob, …) of a specific type (one of the three described above). Only **Role** is strictly required because it dictates the rules that allow other agents to determine what sorts of attestations to look for, and how to interpret them as a history. The only truly relevant **ID Dat**a is the agent's public key, however, which must be part of the record for signature checking in any event.



NFT Record

role

$\mathrm{ID\_data}(\mathrm{Optional})$

$\mathrm{metadata}(\mathrm{Optional})$

signature

$\mathrm{pub\_key}^{x}$

Identity NFT Datagram

The green elements are concatenated, hashed, and signed.[9]

$$\mathrm{Hash}(\mathrm{role}|\mathrm{ID\_data}|\mathrm{metadata}) \equiv \mathrm{hash\_digest}$$

[9] Note that " | " indicates that the byte strings in the argument are **concatenated**.

$$\mathrm{Signature}(\mathrm{pri\_key}^{x},\ \mathrm{hash\_digest}) = \mathrm{signature}\,.$$

It will not matter if an individual Mechanical (whatever that might mean) creates multiple identities. If it does, it is effectively setting-up subsidiaries and “doing business as” several public keys. Since public keys are evaluated on the basis of their own histories, this is no different from separate Mechanicals setting up to do business separately under these public keys. The incentives are the same.

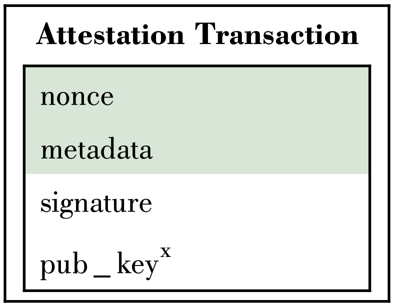
It also will not matter if a Mechanical hands over its private key to another Mechanical. The incentives for the new owner are the same as for the old owner. Behaving honestly has the same expected value no matter who owns the key, and giving away a key is just like replacing the management of a business.

What will matter is if a key-holder knows, or believes that there is a probability, that it will leave the game, or that the game will end. If there is a known final period, then cooperation unravels in the usual way. If the personal or general final period is probabilistic, then periodic payoffs to the Mechanicals must go up commensurately to account for the lessened value of the future. A similar dynamic occurs if the overall market size changes over time. If it is expected to grow, then the value of the future is higher, all else equal, and if it is expected to shrink, it is lower.

Creating multiple identity NFTs with the same public key should be considered *per se* dishonest, and is easily detectable.

9.3. Messaging using Attestation Transactions

Attestation transactions are created and signed by coin account holders on the blockchain. Unlike NFTs, they do not create records. A valid attestation transaction is simply added to current block. The only record it modifies is the sending coin record, which has the required transaction fee deducted, and its nonce incremented.



Attestation Transaction

nonce

metadata

signature

$\mathrm{pub\_key}^{x}$

Attestation Transaction Datagram

Again, the green elements are concatenated, hashed, and signed.

For our purposes here, the identity NFT creation, and all associated attestation transactions, must originate from the same coin account controlled by the private key, $\mathrm{pri\_key}^{x}$ that signs them all. In fact, this can be done much more elegantly, but these details do not change the logic of the architecture.

The metadata elements in the attestation transaction are actually messages of different types that mediate the market and generate provable histories. In the following subsections, we describe the content of the these metadata elements for each of the three agent types.

9.3.1. Biological Message Metadata Content

A Biological $b \in \mathcal{B}$, begins by choosing a Mechanical, $m \in \mathcal{M}$, a Verifier, $v \in \mathcal{V}$, a process identifier, $p \in \mathcal{P}$, and an offer (Fee, p), then creating and committing to the blockchain an **Offer Message** attestation transaction with the following metadata:

Offer Metadata

Offer Message

$\mathrm{ID\_data}^{b}$

$\mathrm{proc\_ID}$

$\mathrm{ID\_data}^{m}/\mathrm{pub\_key}^{m}$

$\mathrm{ID\_data}^{v}/\mathrm{pub\_key}^{v}$

Fee

p

where:

- Offer Message: A plaintext message type label.
- $\mathrm{ID\_data}^{b}$: The ID number chosen by the Biological when creating its identity NFT. This is not strictly necessary since the transaction includes the Biological's public key, which unambiguously identifies the message's originator.
- $\mathrm{proc\_ID}$: $p \in \mathcal{P}$, the process the Biological wishes to have executed.
- $\mathrm{ID\_data}^{m}/\mathrm{pub\_key}^{m}$: The ID number and/or public key of the Mechanical the Biological has chosen. At least one is needed, but the public key makes look-ups easier.
- $\mathrm{ID\_data}^{v}/\mathrm{pub\_key}^{v}$: The ID number and/or public key of the Verifier the Biological has chosen.
- Fee: The fee being offered to the Mechanical.
- p: The probability of audit the Biological will pay for.

Suppose that the Biological commits an offer message that gets included in a block at height N. Suppose for the moment that Mechanical sees this message and responds with an accept message (see the next Subsection). Then the Biological commits an **Input Message** to the blockchain.

Input Metadata

Input Message

Hash(Offer_TX)

$\mathrm{ID\_data}^{m}/\mathrm{pub\_key}^{m}$

$\mathrm{ID\_data}^{v}/\mathrm{pub\_key}^{v}$

$\mathrm{Encrypt}(\mathrm{pub\_key}^{m}, \mathrm{sym\_key})$

$\mathrm{Encrypt}(\mathrm{sym\_key}, \mathrm{input}_{i})$

$\mathrm{Hash}(\mathrm{Coin\_TX}^{m})$

$\mathrm{Hash}(\mathrm{Coin\_TX}^{v})$

where:

- Input Message: As above.
- Hash(Offer_TX) : The hash of the offer message attestation transaction that initiates the exchange. This is used as an identification number to make it easy for a block explorer to collect all messages subsequently connected to a given offer.
- $\mathrm{ID\_data}^{m}/\mathrm{pub\_key}^{m}$: As above. Not strictly necessary since it can be looked up using Hash(Offer_TX).
- $\mathrm{ID\_data}^{v}/\mathrm{pub\_key}^{v}$: As above, and used by the Verifier to find which messages it should pay attention to.
- $\mathrm{Encrypt}(\mathrm{pub\_key}^{m}, \mathrm{sym\_key})$: The Biological generates a random symmetric key, and encrypts it with the public key of the Mechanical.
- $\mathrm{Encrypt}(\mathrm{sym\_key}, \mathrm{input}_{i})$ The Biological uses this symmetric key to encrypt the inputs it wants to have processed. We discuss the reasons for this approach and alternatives in the last Subsection below.
- $\mathrm{Hash}(\mathrm{Coin\_TX}^{m})$: The Biological commits a separate coin transaction sending Fee to the Mechanical and includes the hash of the transaction to allow verification of this fact.
- $\mathrm{Hash}(\mathrm{Coin\_TX}^{v})$: The Biological does the same thing to send $p \times CV$ to the chosen Verifier.

Suppose that the Biological commits an offer or input message that gets included in a block at height N. Any Mechanical that maintains an identity NFT in the ledger is obliged monitor the blockchain for messages. It does not respond within some set number of blocks, it is considered non-responsive, which is the same as noncooperative[10]. In this event, the Biological commits a **No Response Message** to the blockchain.

No Response Metadata

No Response Message

$\mathrm{Hash}(\text{Offer TX})$

$\mathrm{ID\_data}^{m}/\mathrm{pub\_key}^{m}$

where

- No Response Message: As above.
- $\mathrm{Hash}(\text{Offer TX})$: As above.
- $\mathrm{ID\_data}^{m}/\mathrm{pub\_key}^{m}$: As above.

This message should be seen by the Verifier who will commit a verification message, outlined below.

Finally, suppose that all goes well, and the Mechanical commits an output message, and the public randomization device[11] indicates that an audit is called for. Then the Biological commits an **Audit Message** to the blockchain.

Audit Metadata

Audit Message

$\mathrm{Hash}(\text{Offer TX})$

$\mathrm{Encrypt}(\mathrm{pub\_key}^{v}, \mathrm{sym\_key})$

where

- Audit Message: As above.
- $\mathrm{Hash}(\text{Offer TX})$: As above.
- $\mathrm{Encrypt}(\mathrm{pub\_key}^{v}, \mathrm{sym\_key})$: The same symmetric key that the Biological chose for the input message is encrypted with the Verifier's public key. This allows the Verifier to go to the blockchain, find the input and output messages associated with $\mathrm{Hash}(\text{Offer TX})$, decrypt the ciphertext inputs and outputs that are signed and attested to by the Biological and Mechanical, respectively, and run proc_{p} independently.

9.3.2. Mechanical Message Metadata Content

Each Mechanical, $m \in \mathcal{M}$, monitors the blockchain for messages. When it sees an offer message containing $\mathrm{ID\_data}^{m}/\mathrm{pub\_key}^{m}$ it considers the offer (Fee, p) and the Process ID it con-

[10] There is, in fact, a mechanism that allows agents to declare that they are off-line, and then come back on-line at later block height without removing their identity NFT, and with it, the history they have established. We omit these details for now.

[11] For example, the hash of the concatenation of the offer transaction hash, and the Merkle root of the block committed after the one containing the output message could be used as a seed.

tains, if it finds the offer acceptable, then the Mechanical commits an **Accept Message** to the blockchain.

Accept Metadata

Accept Message

Hash(Offer_TX)

ID_data^{m}/pub_key^{m}

where

- Accept Message: As above.
- Hash(Offer TX): As above.
- ID_data^{m}/pub_key^{m}: As above, and not strictly needed since the public key signing the transaction will also serve.

If the offer is not acceptable then the Mechanical commits a **Decline Message** to the blockchain.

Decline Metadata

Decline Message

Hash(Offer_TX)

ID_data^{m}/pub_key^{m}

where:

- Decline Message: As above.
- Hash(Offer TX): As above.
- ID_data^{m}/pub_key^{m}: As above.

Suppose that the Mechanical accepts, and the Biological, in fact, commits a correct input message. Then the Mechanical decides on CORRECT or MALICIOUS execution, generates an output, and, commits an **Output Message** to the blockchain.

Output Metadata

Output Message

Hash(Offer_TX)

ID_data^{m}/pub_key^{m}

$Encrypt(sym\_key, output_{o})$

where:

- Output Message: As above.

- Hash(Offer_TX): As above.
- ID_data^{m}/pub_key^{m} : As above.
- Encrypt(sym_key, output$_{o}$) : The Mechanical uses the same symmetric key as the Biological in its input message to encrypt the output it generates.

The Biological is required to undertake several actions correctly. If he does not, honest Mechanicals are not able to complete their side of the transaction, and should escape sanction. It may be that Biologicals should be sanctioned or labeled as non-cooperative in this event, but we leave this possibility for the future. There are two possibilities.

First, if Mechanical commits an accept message, but the Biological does not commit an input message before a certain number of blocks have passed, then the Mechanical commits a **No Response Message**,

No Response Metadata

No Response Message

Hash(Offer TX)

ID_data^{b}/pub_key^{b}

where

- No Response Message: As above.
- Hash(Offer TX) : As above.
- ID_data^{b}/pub_key^{b} : As above.

Second, if the Biological commits an input message that is flawed in one or more of the following ways:

- Hash(Coin_TX^{m}) and/or Hash(Coin_TX^{v}) is not actually be committed to the blockchain.
- Hash(Coin_TX^{m}) and/or Hash(Coin_TX^{v}) does not transfer the right fee, or is not to or from the right coin accounts.
- ID_data^{m}/pub_key^{m}, ID_data^{m}/pub_key^{m}, and/or Encrypt(pub_key^{m}, sym_key), are inconsistent with the original offer transaction, Hash(Offer_TX), which is hash referenced in the message.

If so, then the Mechanical commits a **Flawed Input Message**,

Flawed Input Metadata

Flawed Input Message

Hash(Offer TX)

ID_data^{b}/pub_key^{b}

where

- Flawed input Message: As above.

- Hash(Offer TX) : As above.
- ID _data^{b}/pub _key^{b} : As above.

In both cases, the message should be seen by the Verifier who will commit a verification message, outlined below.

9.3.3. Verifier Message Metadata Content

Each Verifier, v ∈ $\mathcal{V}$, monitors the blockchain for certain messages, which it analyzes, and if required, chooses a verification code and then commits a verification message to the blockchain. Specifically:

- No response message from the Biological claiming that the Mechanical has neither accepted not declined: If true, then verification code = **Dishonest Mechanical**. If false, then verification code = **Dishonest Biological**.
- No response message from the Mechanical claiming that the Biological has not committed an input message despite the Mechanical having committed an accept message: If true, then verification code = **Dishonest Biological**. If false, then verification code = **Dishonest Mechanical**.
- Flawed input message from the Mechanical claiming that the input message committed by the Biological does not follow the game's messaging rules. If true, then verification code = **Dishonest Biological**. If false, then verification code = **Dishonest Mechanical**.
- No response message from the Biological claiming that the Mechanical has not committed an output message despite the Biological having committed an input message: If true, then verification code = **Dishonest Mechanical**: If false, then verification code = **Dishonest Biological**.
- An output message. If no audit is called for by the public randomization device, then verification code = **Uncertain**.
- An audit message from the Biological when one is required. In this case, the Verifier conducts an audit and decides on a verification code = **Correct** or **Malicious**.
- Finally, if an audit is called for, but the Biological fails to commit an audit message, the Verifier commits a verification message with verification code = **Dishonest Biological**.

In all cases, it can consult the block explorer to find any data needed to confirm or reject any of these claims or outcomes. When it decides on a verification code, the Verifier commits a **Verification Message** to the blockchain.

Verification Metadata

Verification Message

Hash(Offer _TX)

Verification Code

ID _data^{b}/pub _key^{b}

ID _data^{m}/pub _key^{m}

where

- Verification Message: As above.
- Hash(Offer TX) : As above.
- Verification Code: As just described.
- ID_data^{b}/pub_key^{b} : As above, and not strictly needed, but makes the messaging more transparent.
- ID_data^{m}/pub_key^{m} : As above, and not strictly needed, but makes the messaging more transparent.

Note that if the Biological sends a fake symmetric key in its audit message (or an incorrectly encrypted one) to the Mechanical, or if it encrypts an incorrect or unprocessable input, the Mechanical will return whatever garbage output results. The Biological will then be the party that the Verifier identifies as responsible in the event of an audit.

Also note that the Biological cannot send a symmetric key that is different from the one he used in this input message to the Mechanical. The Verifier generates a plaintext of the symmetric key from the encrypted one sent in the audit message. It then only needs to encrypt this with the Mechanical's public key to determine whether the Biological sent the same one as in its input message. Thus, the Verifier will have the same symmetric key used by the Biological and Mechanical in their exchange of messages. The Verifier will therefore end up with the same plaintext inputs and outputs and the two parties, and will be able to verify whether the Mechanical behaved honestly.

9.3.4. A Summary of Message Flow

The Message Flow Table below shows the order of messages along all the possible paths, which depend on the actions taken by the three agents. The subscripts indicate the block height at which a message was committed. The cells shaded green show paths and outcomes in which all agents sent and responded to messages within the game's messaging rules. The cells shaded in red show paths and outcomes where one of the agents did not send messages as required the game's rules, and which result in a verifier message assigning responsibility.

<table>
<tr><th colspan="8"><b>Message Flow Table</b></th></tr>
<tr><td colspan="8">OFF^{b}_{N0}</td></tr>
<tr><td colspan="6">ACC^{m}_{N1}</td><td>DCL^{m}_{N1}</td><td>NR^{b}_{N1}</td></tr>
<tr><td colspan="5">INP^{b}_{N2}</td><td>NR^{m}_{N2}</td><td rowspan="2">$\mathrm{VM}^{v}_{N2} =$ NUL</td><td rowspan="2">$\mathrm{VM}^{v}_{N2} =$ DB/DM</td></tr>
<tr><td colspan="3">OUT^{m}_{N3}</td><td>FI^{m}_{N3}</td><td>NR^{b}_{N3}</td><td rowspan="2">$\mathrm{VM}^{m}_{N3} =$ DB/DM</td></tr>
<tr><td colspan="2">$(1-p)$</td><td>p</td><td rowspan="2">$\mathrm{VM}^{m}_{N4} =$ DB/DM</td><td rowspan="2">$\mathrm{VM}^{m}_{N4} =$ DB/DM</td><td></td><td></td></tr>
<tr><td>AUD^{b}_{N4}</td><td rowspan="2">$\mathrm{VM}^{v}_{N4} =$ DB</td><td rowspan="2">$\mathrm{VM}^{m}_{N4} =$ UNC</td><td></td><td></td><td></td></tr>
<tr><td rowspan="2">$\mathrm{VM}^{v}_{N5} =$ COR/MAL</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
</table>

The Legend and Details for Message Flow Table provides some detail and context for the first table. The main new element is the Creation Time Limits column. Once a Biological commits an offer messages to a block at height N0, other agents must respond withing certain time intervals.

The Mechanical is required to commit an accept or decline message before a limit of **L** additional blocks have been committed to the chain (that is, before some block $N1 < NO + L$). In the event that an accept message in committed at block N1, the Biological is required to commit an input message at some block $N2 < N1 + L$. In all cases where a response is needed from a specific agent, the game's messaging rules require that it be committed before the block limit expires or else the agent is deemed to be non-responsive, and therefore dishonest.

On the other hand, no response claims by Biologicals and Mechanicals cannot be committed before the block limited expires ($N2 > N1 + L$, for example), and need not be committed at all. If a no response message is committed, then the Verifier is required to commit a verification message within the normal block limit ($N3 < N2 + L$, for example). In the case where an audit is called for but the Biological fails to commit an audit message containing the key, both limits apply. That is, the Verifier must wait until the block limit for committing the audit message has expired, but then must commit its verification message within its own block limit, $N4 \in (N3+L, N3+2L)$.

Legend and Details for Message Flow Table

Symbol	Message Type	Creation Time Limits	Key Content or Purpose
OFF^{b}_{N0}	Offer Message	$N0 = $ Initial time	$\mathrm{proc\_ID},\ m,\ v,\ (\mathrm{Fee}, p)$
ACC^{m}_{N1}	Accept Message	$N1 < N0+L$	Accepted offer, send input
DCL^{m}_{N1}	Decline Message	$N1 < N0+L$	Declined offer, NULL execution
NR^{b}_{N1}	No Response Message	$N1 > N0+L$	No ACC^{m} or DLC^{m} received
INP^{b}_{N2}	Input Message	$N2 < N1+L$	$\mathrm{sym\_key},\ \mathrm{input}_{i}$
NR^{m}_{N2}	No Response Message	$N2 < N1+L$	No INP^{m} received
VM^{v}_{N2}	Verifier Message	$N2 < N1+L$	NUL event
VM^{v}_{N2}	Verifier Message	$N2 < N1+L$	Dishonest Bio or Mech (No ACC^{m} or DLC^{m} received)
OUT^{m}_{N3}	Output Message	$N3 < N2+L$	output_{o}
FI^{m}_{N2}	Flawed Input Message	$N2 < N1+L$	Flawed input message
NR^{b}_{N3}	No Response Message	$N3 > N2+L$	No OUT^{m} received
VM^{m}_{N3}	Verifier Message	$N3 < N2+L$	Dishonest Bio or Mech (No INP^{b} received)
AUD^{b}_{N4}	Audit Message	$N4 < N3+L$	$\mathrm{sym\_key}$
VM^{m}_{N4}	Verifier Message	$N4 < N4+L$	Dishonest Bio (No AUD^{b} received)
VM^{m}_{N4}	Verifier Message	$N4 < N3+L$	UNC event
VM^{v}_{N4}	Verifier Message	$N4 \in (N3+L, N3+2L)$	Dishonest Bio
VM^{m}_{N4}	Verifier Message	$N4 < N3+L$	Dishonest Bio or Mech (Flawed input message)
VM^{m}_{N4}	Verifier Message	$N4 < N3+L$	Dishonest Bio or Mech (No OUT^{m} received)
VM^{v}_{N5}	Verifier Message	$N5 < N4+L$	COR or MAL event

9.3.5. A Less Data Intensive Approach

Above, we described a robust, but informationally costly, approach to input, output, and audit messages. Specifically, the input and output messages contain the full ciphertext of the literal inputs and outputs. This makes it impossible for either the Biological or the Mechanical to deny what was sent or received, and allows the Verifier to determine the type of execution the Mechanical chose using only the relevant symmetric key.

If we are willing to allow more rounds of communication, then we can reduce the data burden of attestation transactions as follows:

- The Biological replaces $\mathrm{Encrypt}(\mathrm{pub\_key}^{m}, \mathrm{sym\_key})$ and $\mathrm{Encrypt}(\mathrm{sym\_key}, \mathrm{input}_{i})$ in the input message with $\mathrm{Hash}(\mathrm{input}_{i})$.
- If the Mechanical accepts, the Biological sends the Mechanical the full text of the input out-of-band.
- The Mechanical must then either commit an acknowledgment message that includes the hash of input to confirm what he received, or a no response message claiming the either it never got the input, or that it was different from the hash in the input message.
- In the event of a no response message from the Mechanical, the Biological must commit a new input message with the full ciphertext of the input.
- Things proceed as before until the Mechanical is ready to send its output. The pattern above is followed.
- The Mechanical replaces $\mathrm{Encrypt}(\mathrm{sym\_key}, \mathrm{output}_{o})$ with $\mathrm{Hash}(\mathrm{output}_{o})$ in its output message and then sends the Biological the full text of the output out-of-band.
- The Biological must then either commit an acknowledgment message that includes the hash of its output to confirm what he received, or a no response message claiming the either he never got the output, or that it was different from the hash in the output message.
- In the event of a no response message from the Biological, the Mechanical must commit a new output message with the full ciphertext of the output.
- If an audit is called for at this point, the Biological has both the input and output that were either hashed, or encrypted, and then committed to a block. If only the hashes are in the messages, the Biological is required to send the plaintext of both to the verifier out-of-band.
- If they are not committed, the Verifier commits a no response claim, and the Biological must commit the full the ciphertexts to a block or be judged dishonest. Since the signed hashes are in the chain, the Biological cannot send false inputs or outputs.

Note that the blockchain is used as a kind of billboard in the sense that agents cannot pretend to be unaware of messages directed to them. This is key because otherwise it is impossible to differentiate intentional, strategic, silence or deafness, from true communications failure. If data is in the blockchain, it is both provably sent, and provably received, at least within game messaging rules. Consequently, one would hope that in almost all cases, the existence of a mechanism that makes it impossible for agents to deny that they sent or received the full inputs or outputs would make it use rare. Sending full encrypted inputs and outputs through the blockchain is more costly to both parties, and does not produce a strategic advantage for either. Thus, signed hashes will most likely suffice.

10. Mathematical Appendix

The Theorem 1 says that the only SPE equilibrium in the one-shot game is for Biological to choose PASS rather than making an offer to the Mechanical to execute a process. This results in a loss of potential gains from trade due to the non-contractibility of CORRECT process execution.

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

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

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

$$s^b = \mathrm{PASS} \text{ and } s^m(\mathrm{Fee},\, p) = \mathrm{MALICIOUS},$$

and

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

Proof:

Suppose that

$$s^b = \mathrm{PASS}.$$

Then the Mechanical is constrained to choose

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

which is therefore (trivially) a best-response.

Suppose instead that:

$$s^b \neq \bar{s}^b = (\overline{\mathrm{Fee}},\, \bar{p}) \in [0,\, \bar{F}] \times [0,\, 1].$$

Then

$$F^m((\overline{\mathrm{Fee}},\, \bar{p}),\, \mathrm{MALICIOUS}) = \overline{\mathrm{Fee}} + \mathrm{MaliciousValue}(\mathrm{input}_i) >$$
$$F^m((\overline{\mathrm{Fee}},\, \bar{p}),\, \mathrm{CORRECT}) = \overline{\mathrm{Fee}} - \mathrm{CostProc}(\mathrm{Proc}_p),$$

and

$$F^m((\overline{\mathrm{Fee}},\, \bar{p}),\, \mathrm{MALICIOUS}) = \overline{\mathrm{Fee}} + \mathrm{MaliciousValue}(\mathrm{input}_i) >$$
$$F^m((\overline{\mathrm{Fee}},\, \bar{p}),\, \mathrm{NULL}) = 0,$$

and so the Mechanical will always choose

$$s^m((\overline{\mathrm{Fee}},\, \bar{p})) = \mathrm{MALICIOUS}$$

in the subgames where $\bar{s}^b = (\overline{\mathrm{Fee}},\, \bar{p})$.

Since

$$F^b(\mathrm{PASS},\, \mathrm{MALICIOUS}) = 0 >$$
$$F^b((\overline{\mathrm{Fee}},\, \bar{p})),\, \mathrm{MALICIOUS}) = -\overline{\mathrm{Fee}} - \bar{p} \times \mathrm{CostVerify}(\mathrm{Proc}_p) - \varepsilon$$

The Biological will therefore always prefer the subgame where he chooses:

$$s^b = \mathrm{PASS}.$$

■

Lemma 1 says that in any period T where the history is noncooperative, playing the strategy profile $S_{\infty} = (\mathrm{Grim}_{\infty}, \mathrm{MinAccept}_{\infty})$ gives the maximal period T payoffs to each agent.

Lemma 1:

$$\forall\ T \in \mathcal{T} \text{ and } \forall\ H_T \notin \mathcal{H}^{\mathrm{coop}},$$

if

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

then

$$\forall\ \bar{S}_{\infty}^{b} \in \mathcal{S}_{\infty}^{b} \text{ and } \forall\ \bar{S}_{\infty}^{m} \in \mathcal{S}_{\infty}^{m},$$

it holds that

$$F^{b}(s_T^{b}(H_T),\ s_T^{m}(H_T,\ s_T^{b}(H_T))) = 0 \geq\ F^{b}(\bar{s}_T^{b}(H_T),\ s_T^{m}(H_T,\ \bar{s}_T^{b}(H_T)))$$

and

$$F^{m}(\bar{s}_T^{b}(H_T),\ s_T^{m}(H_T,\ \bar{s}_T^{b}(H_T))) = 0 \geq\ F^{m}(\bar{s}_T^{b}(H_T),\ \bar{s}_T^{m}(H_T,\ \bar{s}_T^{b}(H_T))).$$

Proof:

(A) First, consider the Biological.

If

$$\mathrm{Grim}_T^{b}(H_T) \neq \bar{s}_T^{b}(H_T) = \bar{a}_T^{b} = (\overline{\mathrm{Fee}},\ \bar{p}) \in [0,\ \bar{F}]\times[0,\ 1],$$

then

$$\mathrm{MinAccept}_{\infty}(H_T,\ \bar{a}_b^{T}) = s_T^{m}(H_T,\ \bar{a}_b^{T}) = a_T^{m} = \mathrm{MALICIOUS}$$

$$F^{b}((\overline{\mathrm{Fee}},\ \bar{p}),\ \mathrm{MALICIOUS}) = -\ \overline{\mathrm{Fee}}\ -\ \bar{p}\times \mathrm{CV}\ -\ \varepsilon < 0,$$

and if

$$\mathrm{Grim}_T^{b}(H_T) = s_T^{b}(H_T) = a_T^{b} = \mathrm{PASS},$$

then

$$\mathrm{MinAccept}_{\infty}(H_T,\ a_T^{b}) = s_T^{m}(H_T,\ a_T^{b}) = a_T^{m} = \mathrm{NULL}$$

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

Thus,

$$\forall\ \bar{S}_{\infty}^{b} \in \mathcal{S}_{\infty}^{b}$$

it holds that

$$F^{b}(\mathrm{Grim}_T(H_T),\ \mathrm{MinAccept}_T(H_T,\ \mathrm{Grim}_T(H_T))) = 0 \geq F^{b}(\bar{s}_T^{b}(H_T),\ \mathrm{MinAccept}_T(H_T,\ \bar{s}_T^{b}(H_T))).$$

(B) Next, consider the Mechanical.

If

$$\mathrm{Grim}_T^{b}(H_T) \neq \bar{s}_T^{b}(H_T) = \bar{a}_T^{b} = (\overline{\mathrm{Fee}},\ \bar{p}) \in [0,\ \bar{F}]\times[0,\ 1],$$

then

$$\mathrm{MinAccept}_T(H_T,\ \bar{a}_T^{b}) = \mathrm{MALICIOUS}$$

and

$$F^{m}((\overline{\mathrm{Fee}},\ \bar{p}), \mathrm{MALICIOUS}) = \overline{\mathrm{Fee}} + \mathrm{MV} >\ F^{m}((\overline{\mathrm{Fee}},\ \bar{p}),\ \mathrm{CORRECT}) = \overline{\mathrm{Fee}} - \mathrm{CP},$$

$$F^{m}((\overline{Fee},\bar{p}),\mathrm{MALICIOUS})=\overline{Fee}+MV>F^{m}((\overline{Fee},\bar{p}),\mathrm{NULL})=0,$$

and if,

$$\mathrm{Grim}^{b}_{T}(H_T)=s^{b}_{T}(H_T)=a^{b}_{T}=\mathrm{PASS},$$

then NULL is the only choice available to the Mechanical, and

$$\mathrm{MinAccept}_T(H_T,\bar{a}^{b}_{T})=\mathrm{NULL}$$

and the Mechanical must choose NULL

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

Thus,

$$\forall\ \bar{S}^{b}_{\infty}\in\mathcal{S}^{b}_{\infty}\ \text{ and }\ \forall\ \bar{S}^{m}_{\infty}\in\mathcal{S}^{m}_{\infty},$$

it holds that

$$F^{m}(\bar{s}_T(H_T),\mathrm{MinAccept}_T(H_T,\bar{s}^{b}_{T}(H_T)))=0\geq F^{m}(\bar{s}_T(H_T),\bar{s}^{m}_{T}(H_T,\bar{s}^{b}_{T}(H_T))).$$

■

The Lemma 2 says that in any period T where the history is noncooperative, playing the strategy profile $S_\infty=(\mathrm{Grim}_\infty,\ \mathrm{MinAccept}_\infty)$ gives the maximal expected payoffs in the continuation game to each agent.

Lemma 2:

$$\forall\ T\in\mathcal{T}\ \text{ and }\ \forall\ H_T\notin\mathcal{H}^{coop},$$

if

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

then

$$\forall\ \bar{S}^{b}_{\infty}\in\mathcal{S}\infty^{b}\ \text{ and }\ \forall\ \bar{S}^{m}_{\infty}\in\mathcal{S}^{m}_{\infty}$$

it holds that

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

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

and

$$EPO^{m}(0,H_0,B^{b}_{\infty},B^{m}_{\infty})=\mathrm{MaxEPO}^{m}(0,H_0,B^{b}_{\infty}).$$

Proof:

(A) First, consider the Biological.

If

$$H_T\notin\mathcal{H}^{coop},$$

then by Lemma 1,

$$\forall\ \bar{S}^{b}_{\infty}\in\mathcal{S}^{b}_{\infty}$$

$$F^{b}(\mathrm{Grim}_T(H_T),\mathrm{MinAccept}_T(H_T,\mathrm{Grim}_T(H_T))=0\geq F^{b}(\bar{s}^{b}_{T}(H_T),\mathrm{MinAccept}_T(H_T,\bar{s}^{b}_{T}(H_T))).$$

and since if

$$H_T\notin\mathcal{H}^{coop},$$

then

$$\forall\ t > T,\ H_t \notin \mathcal{H}^{coop},$$

and the inequality continues to hold for future periods, which implies that $\mathrm{Grim}_t(H_t)$ gives the Biological the highest possible periodic payoff when the Mechanical chooses strategy $\mathrm{MinAccept}_t(H_t)$. It follows that:

$$\forall\ \bar{S}^b_\infty \in \mathcal{S}^b_\infty$$

$$\mathrm{EPO}^b(T, H_T, \mathrm{Grim}_\infty, \mathrm{MinAccept}_\infty) \geq \mathrm{EPO}^b(T, H_T, \bar{S}^b_\infty, \mathrm{MinAccept}_\infty).$$

(B) Next, consider the Mechanical.

If

$$H_T \notin \mathcal{H}^{coop},$$

then by Lemma 1,

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

it holds that

$$F^m(\bar{s}^b_T(H_T), \mathrm{MinAccept}_T(H_T, \bar{s}^b_T(H_T))) = 0 \geq F^m(\bar{s}^b_T(H_T), \bar{s}^m_T(H_T, \bar{s}^b_T(H_T))).$$

and since, as above, if

$$H_T \notin \mathcal{H}^{coop},$$

then it is also the case that

$$\forall\ t > T,\ H_t \notin \mathcal{H}^{coop},$$

and the inequality continues to hold for future periods, which implies that $\mathrm{MinAccept}_t(H_t)$ gives the Mechanical the highest possible periodic payoff regardless of the strategy Biological chooses. It follows that:

$$\forall\ \bar{S}^m_\infty \in \mathcal{S}^m_\infty$$

$$\mathrm{EPO}^m(T, H_T, \mathrm{Grim}_\infty, \mathrm{MinAccept}_\infty) \geq \mathrm{EPO}^m(T, H_T, \mathrm{Grim}_\infty, \bar{S}^b_\infty),$$

and since this also holds for

$$T = 0,\ H_0 \notin \mathcal{H}^{coop},$$

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

■

The Lemma 3 says that in any period T where the history is cooperative, playing the strategy Grim_∞ when the Mechanical plays $\mathrm{MinAccept}_\infty$ gives the maximal the period T payoff to the Biological.

Lemma 3:

$$\forall\ T \in \mathcal{T} \text{ and } \forall\ H_T \in \mathcal{H}^{coop},$$

if

$$S^b_\infty = \mathrm{Grim}_\infty \text{ and } S^m_\infty = \mathrm{MinAccept}_\infty,$$

then

$$\forall\ \bar{S}^b_\infty \in \mathcal{S}^b_\infty$$

it holds that

$$F^{b}(s_T^{b}(H_T),\ s_T^{m}(H_T,\ s_T^{b}(H_T)) = 0 > \ F^{b}(\bar{s}_T^{b}(H_T),\ s_T^{m}(H_T,\ \bar{s}_T^{b}(H_T)).$$

Proof:

(A) Suppose for some $T \in \mathcal{T}$,

$$\exists\ a_T^{b} = (Fee,\ p) = \underset{(Fee,\ p) \in [0,\ \bar{F}] \times [0,\ 1]}{\text{argmin}} \quad Fee + p \times CV$$

such that

$$EPO_C(Fee,\ p) \geq EPO_D(Fee,\ p)$$
$$U - (Fee + p \times CV) \geq 0.$$

(a) Suppose first that,

$$Grim_T(H_T) \neq \bar{s}_T^{b}(H_T) = \bar{a}_t^{b} = (\overline{Fee},\ \bar{p}),\ \neq (Fee,\ p),$$

and

$$\frac{\overline{Fee} - CP}{(1 - r)} \equiv EPO_C(\overline{Fee},\ \bar{p}) > \ EPO_D(\overline{Fee},\ \bar{p}) \equiv \frac{\overline{Fee} + MV}{(1 - r + r\bar{p})}.$$

Then

$$MinAccept_T(H_T,\ \bar{a}_T^{b}) = CORRECT.$$

However, for some

$$\tilde{a}_T^{b} = (\widetilde{Fee},\ \bar{p})\ \text{where}\ \widetilde{Fee} < \overline{Fee},$$

this inequality continues to hold, and

$$F^{b}((\widetilde{Fee},\ \bar{p}),\ CORRECT) = U - \widetilde{Fee} - \bar{p} \times CV >$$
$$F^{b}((\overline{Fee},\ \bar{p}),\ CORRECT) = U - \overline{Fee} - \bar{p} \times CV \geq 0.$$

Thus, if

$$\bar{a}_t^{b} = (\overline{Fee},\ \bar{p}),\ \neq (Fee,\ p)$$

then it cannot be the case that this is a period T payoff maximizing action.

(b) Suppose second that

$$Grim_T(H_T) = s_T^{b}(H_T) = a_t^{b} = (Fee,\ p)$$

and

$$EPO_C(Fee,\ p) = EPO_D(Fee,\ p).$$

Then

$$MinAccept_T(H_T,\ a_T^{b}) = CORRECT$$
$$F^{b}((Fee,\ p),\ CORRECT) = U - Fee - p \times CV \geq 0.$$

(c) Suppose third that,

$$Grim_T(H_T) \neq \bar{s}_T^{b}(H_T) = \bar{a}_t^{b} = (\overline{Fee},\ \bar{p}),\ \neq (Fee,\ p)$$

and

$$EPO_C(\overline{Fee},\ \bar{p}) < EPO_D(\overline{Fee},\ \bar{p}).$$

Then

$$\text{MinAccept}_T(H_T, \bar{a}^b_T) = \text{MALICIOUS}$$
$$F^b((\overline{\text{Fee}}, \bar{p}), \text{MALICIOUS}) = -\overline{\text{Fee}} - \bar{p}\times CV - \varepsilon < 0.$$

(d) Suppose fourth that,

$$\text{Grim}_T(H_T) \neq \bar{s}^b_T(H_T) = \bar{a}^b_t = \overline{\text{PASS}}.$$

Then

$$\text{MinAccept}_T(H_T, \bar{a}^b_T) = \text{NULL}$$
$$F^b(\overline{\text{PASS}}, \text{NULL}) = 0.$$

(B) Suppose instead that for some $T \in \mathcal{T}$,

$$\nexists\ a^b_T = (\text{Fee}, p) = \underset{(\text{Fee},\, p) \in [0,\, \bar{F}]\times[0,\, 1]}{\text{argmin}} \quad \text{Fee} + p\times CV$$

such that

$$EPO_C(\text{Fee}, p) \geq EPO_D(\text{Fee}, p)$$
$$U - (\text{Fee} + p\times CV) \geq 0.$$

(a) Suppose first that,

$$\text{Grim}_T(H_T) \neq \bar{s}^b_T(H_T) = \bar{a}^b_t = (\overline{\text{Fee}}, \bar{p})$$

and

$$EPO_C(\overline{\text{Fee}}, \bar{p}) \geq EPO_D(\overline{\text{Fee}}, \bar{p})$$

Then

$$\text{MinAccept}_T(H_T, \bar{a}^b_T) = \text{CORRECT}.$$
$$F^b((\overline{\text{Fee}}, \bar{p}), \text{CORRECT}) = U - (\overline{\text{Fee}} + \bar{p}\times CV) < 0.$$

(b) Suppose second that,

$$\text{Grim}_T(H_T) \neq \bar{s}^b_T(H_T) = \bar{a}^b_t = (\overline{\text{Fee}}, \bar{p})$$

and

$$EPO_C(\overline{\text{Fee}}, \bar{p}) < EPO_D(\overline{\text{Fee}}, \bar{p}).$$

Then,

$$s^m(H_T, a^b_T) = \text{MALICIOUS}$$
$$F^b((\overline{\text{Fee}}, \bar{p}), \text{MALICIOUS}) = (\text{Fee} + p\times CV) - \varepsilon < 0.$$

(c) Suppose third that,

$$\text{Grim}_T(H_T) = s^b_T(H_T) = a^b_t = \text{PASS}.$$

Then

$$\text{MinAccept}_T(H_T, \bar{a}^b_T) = \text{NULL}$$
$$F^b(\text{PASS}, \text{NULL}) = 0.$$

Thus, regardless of whether the period T history is cooperative or noncooperative, we conclude:

$$\forall\ T \in \mathcal{T} \text{ and } \forall\ H_T \in \mathcal{H}^{coop},$$

if

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

then

$$\forall\ \bar{S}^{b}_{\infty} \in \mathcal{S}^{b}_{\infty}$$

it holds that

$$F^{b}(s^{b}_{T}(H_{T}),\ s^{m}_{T}(H_{T},\ s^{b}_{T}(H_{T}))) = 0 > F^{b}(\bar{s}^{b}_{T}(H_{T}),\ s^{m}_{T}(H_{T},\ \bar{s}^{b}_{T}(H_{T}))).$$

■

The Lemma 4 says that in any period T where the history is cooperative, playing the strategy Grim_{∞} when the Mechanical plays $\mathrm{MinAccept}_{\infty}$ gives maximal expected payoff in the continuation game to the Biological.

Lemma 4:

$$\forall\ T \in \mathcal{T}, \text{ and } \forall\ H_{T} \in \mathcal{H}^{\mathrm{coop}},$$

if

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

then

$$\forall\ \bar{S}^{b}_{\infty} \in \mathcal{S}^{b}_{\infty}$$

it holds that

$$\mathrm{EPO}^{b}(T,\ H_{T},\ S^{b}_{\infty},\ B^{m}_{\infty}) \geq \mathrm{EPO}^{b}(T,\ H_{T},\ \bar{S}^{b}_{\infty},\ B^{m}_{\infty}).$$

Proof:

Suppose for some $T \in \mathcal{T}$,

$$\bar{s}^{b}_{T}(H_{T}) \neq \mathrm{Grim}_{T}(H_{T})$$

and it happens that

$$H_{(T+1)} \in \mathcal{H}^{\mathrm{coop}}_{\infty},$$

and further suppose,

$$F^{b}(\mathrm{Grim}_{T},\ \mathrm{MinAccept}_{T}) + r \times \mathrm{EPO}^{b}(T+1,\ H_{(T+1)},\ \mathrm{Grim}_{\infty},\ \mathrm{MinAccept}_{\infty}) <$$
$$F^{b}(\bar{s}^{b}_{T},\ \mathrm{MinAccept}_{T}) + r \times \mathrm{EPO}^{b}(T+1,\ H_{(T+1)},\ \bar{S}^{b}_{\infty},\ \mathrm{MinAccept}_{\infty}),$$

which is equivalent to the contradiction of the Lemma's statement.

But by Lemma 3,

$$F^{b}(\mathrm{Grim}_{T}(H_{T}),\ \mathrm{MinAccept}_{T}(H_{T},\ \mathrm{Grim}_{T}(H_{T}))) \geq$$
$$F^{b}(\bar{s}^{b}_{T}(H_{T}),\ \mathrm{MinAccept}_{T}(H_{T},\ \mathrm{Grim}_{T}(H_{T}))),$$

which implies that it must be the case that:

$$\mathrm{EPO}^{b}(T+1,\ H_{(T+1)},\ \mathrm{Grim}_{\infty},\ \mathrm{MinAccept}_{\infty}) <$$
$$\mathrm{EPO}^{b}(T+1,\ H_{(T+1)},\ \bar{s}^{b}_{\infty},\ \mathrm{MinAccept}_{\infty}).$$

By the same argument, this inequality must also hold for all future periods $t > T$ such that $H_{t} \in \mathcal{H}^{\mathrm{coop}}_{\infty}$. But this can only be true if for at least some future period, $\bar{T} > T$,

$$\mathrm{F}^{\mathrm{b}}(\mathrm{Grim}_{\bar{\mathrm{T}}}(\mathrm{H}_{\bar{\mathrm{T}}}), \mathrm{MinAccept}_{\bar{\mathrm{T}}}(\mathrm{H}_{\bar{\mathrm{T}}}, \mathrm{Grim}_{\bar{\mathrm{T}}}(\mathrm{H}_{\bar{\mathrm{T}}}))) \geq \mathrm{F}^{\mathrm{b}}(\bar{\mathrm{s}}^{\mathrm{b}}_{\bar{\mathrm{T}}}(\mathrm{H}_{\bar{\mathrm{T}}}), \mathrm{MinAccept}_{\bar{\mathrm{T}}}(\mathrm{H}_{\mathrm{T}}, \bar{\mathrm{s}}^{\mathrm{b}}_{\bar{\mathrm{T}}}(\mathrm{H}_{\bar{\mathrm{T}}})))$$

which contradicts Lemma 3.

Suppose instead that for some future period $\bar{\mathrm{T}} > \mathrm{T}$, $\mathrm{H}_{\bar{\mathrm{T}}} \notin \mathcal{H}^{\mathrm{coop}}_{\infty}$, and suppose the $\bar{\mathrm{T}}$ is the first such period. Then by Lemma 1,

$$\forall\ \bar{\mathrm{T}} \in \mathcal{T} \text{ and } \forall\ \mathrm{H}_{\bar{\mathrm{T}}} \notin {}^{\mathrm{coop}},$$

if

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

then

$$\forall\ \bar{\mathrm{S}}^{\mathrm{b}}_{\infty} \in \mathcal{S}^{\mathrm{b}}_{\infty}$$

it holds that

$$\mathrm{F}^{\mathrm{b}}(\mathrm{Grim}_{\bar{\mathrm{T}}}(\mathrm{H}_{\bar{\mathrm{T}}}), \mathrm{MinAccept}_{\bar{\mathrm{T}}}(\mathrm{H}_{\bar{\mathrm{T}}}, \mathrm{Grim}_{\bar{\mathrm{T}}}(\mathrm{H}_{\bar{\mathrm{T}}}))) = 0 \geq \mathrm{F}^{\mathrm{b}}(\bar{\mathrm{s}}^{\mathrm{b}}_{\bar{\mathrm{T}}}(\mathrm{H}_{\bar{\mathrm{T}}}), \mathrm{MinAccept}_{\bar{\mathrm{T}}}(\mathrm{H}_{\bar{\mathrm{T}}}, \bar{\mathrm{s}}^{\mathrm{b}}_{\bar{\mathrm{T}}}(\mathrm{H}_{\bar{\mathrm{T}}})).$$

Thus,

$$\mathrm{EPO}^{\mathrm{b}}(\bar{\mathrm{T}}, \mathrm{H}_{\bar{\mathrm{T}}}, \bar{\mathrm{s}}^{\mathrm{b}}_{\infty}, \mathrm{MinAccept}_{\infty}) \leq 0 = \mathrm{EPO}^{\mathrm{b}}(\bar{\mathrm{T}}, \mathrm{H}_{\bar{\mathrm{T}}}, \mathrm{Grim}_{\infty}, \mathrm{MinAccept}_{\infty}).$$

Since for all periods $\mathrm{t} \in (\mathrm{T}, \bar{\mathrm{T}} - 1)$ where $\mathrm{H}_{\mathrm{t}} \notin \mathcal{H}^{\mathrm{coop}}_{\infty}$, we have already established that,

$$\mathrm{F}^{\mathrm{b}}(\mathrm{Grim}_{\mathrm{T}}(\mathrm{H}_{\bar{\mathrm{T}}}), \mathrm{MinAccept}_{\bar{\mathrm{T}}}(\mathrm{H}_{\bar{\mathrm{T}}}, \mathrm{Grim}_{\bar{\mathrm{T}}}(\mathrm{H}_{\bar{\mathrm{T}}}))) \geq \mathrm{F}^{\mathrm{b}}(\bar{\mathrm{s}}^{\mathrm{b}}_{\bar{\mathrm{T}}}(\mathrm{H}_{\bar{\mathrm{T}}}), \mathrm{MinAccept}_{\bar{\mathrm{T}}}(\mathrm{H}_{\mathrm{T}}, \bar{\mathrm{s}}^{\mathrm{b}}_{\bar{\mathrm{T}}}(\mathrm{H}_{\bar{\mathrm{T}}})))$$

we conclude that,

$$\mathrm{EPO}^{\mathrm{b}}(\mathrm{T}, \mathrm{H}_{\mathrm{T}}, \mathrm{Grim}_{\infty}, \mathrm{MinAccept}_{\infty}) \geq \mathrm{EPO}^{\mathrm{b}}(\mathrm{T}, \mathrm{H}_{\mathrm{T}}, \bar{\mathrm{s}}^{\mathrm{b}}_{\infty}, \mathrm{MinAccept}_{\infty}),$$

which proves the Lemma.

■

The Lemma 5 says that in any period T where the history is cooperative and agents play the strategy profile $\mathrm{S}_{\infty} = (\mathrm{Grim}_{\infty}, \mathrm{MinAccept}_{\infty})$, the value of the continuation game for the Mechanical must equal either the expected payoff of choosing CORRECT execution in each period, or of choosing MALICIOUS execution in every period in which the Biological makes an offer instead of choosing PASS.

Lemma 5:

$$\forall\ \mathrm{T} \in \mathcal{T} \text{ and } \forall\ \mathrm{H}_{\mathrm{T}} \in \mathcal{H}^{\mathrm{coop}},$$

if

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

then either

$$\mathrm{MaxEPO}^{\mathrm{m}}(\mathrm{T}, \mathrm{H}_{\mathrm{T}}, \mathrm{B}^{\mathrm{b}}_{\infty}) = \mathrm{EPO}_{\mathrm{C}}(\mathrm{Fee}, \mathrm{p})$$

or

$$\mathrm{MaxEPO}^{\mathrm{m}}(\mathrm{T}, \mathrm{H}_{\mathrm{T}}, \mathrm{B}^{\mathrm{b}}_{\infty}) = \mathrm{EPO}_{\mathrm{D}}(\mathrm{Fee}, \mathrm{p}).$$

Proof:

(A) Suppose for some $T \in \mathcal{T}$,

$$\mathrm{Grim}_{\infty}(H_T) = \beta_T(H_T) = a_T^b = (\mathrm{Fee}, p) \in [0, \overline{F}] \times [0, 1].$$

The Mechanical must choose CORRECT, MALICIOUS, or NULL execution in response, and so at least one of the following must be true, respectively:

(a)

$$\mathrm{MaxEPO}^m(T, H_T, \mathrm{Grim}_{\infty}) = C + r \times \mathrm{MaxEPO}^m(T+1, H_{(T+1)}, \mathrm{Grim}_{\infty})$$

where

$$H_{(T+1)} \in \mathcal{H}_{\infty}^{coop},$$

or

(b)

$$\mathrm{MaxEPO}^m(T, H_T, \mathrm{Grim}_{\infty}) = D + (1 - p) r \times \mathrm{MaxEPO}^m(T+1, H_{(T+1)}, \mathrm{Grim}_{\infty}) > 0$$

where

$$H_{(T+1)} \in \mathcal{H}_{\infty}^{coop},$$

or

(c)

$$\mathrm{MaxEPO}^m(T, H_T, \mathrm{Grim}_{\infty}) = 0 + r \times \mathrm{MaxEPO}^m(T+1, H_{(T+1)}, \mathrm{Grim}_{\infty}) = 0$$

where

$$H_{(T+1)} \notin \mathcal{H}_{\infty}^{coop}.$$

To see this, note the following:

(a) If

$$\mathrm{MinAccept}_T(H_T, a_T^b) = \mathrm{CORRECT},$$

then

$$H_{(T+1)} \in \mathcal{H}_{\infty}^{coop},$$

and so the discounted value of the continuation game is received with certainty.

(b) If

$$\mathrm{MinAccept}_T(H_T, a_T^b) = \mathrm{MALICIOUS},$$

then with probability $(1 - p)$, no audit takes place, $h_{(T+1)} = \mathrm{UNC}$, and $H_{(T+1)} \in \mathcal{H}_{\infty}^{coop}$. With probability p, an audit takes place, $h_{(T+1)} = \mathrm{MAL}$, and $H_{(T+1)} \notin \mathcal{H}_{\infty}^{coop}$. Thus, with probability $(1 - p)$ the Mechanical receives the discounted value of the continuation game with a cooperative history, and with probability p, receives the discounted value of the continuation game with a noncooperative history, and

$$\forall\ t > T, H_t \notin \mathcal{H}_{\infty}^{coop}.$$

To see that the noncooperative continuation game has an expected payoff of zero, note that by Lemma 1:

$$\forall\ T \in \mathcal{T},\ \text{ and } \forall\ H_T \notin \mathcal{H}\{\text{coop}\},$$

if

$$S^b_\infty = \text{Grim}_\infty \ \text{ and } \ S^m_\infty = \text{MinAccept}_\infty$$

then

$$\forall\ \bar{S}^m_\infty \in \mathcal{S}^m_\infty,$$

it holds that

$$F^m(\text{Grim}_T(H_T),\ \text{MinAccept}_T(H_T,\ \text{Grim}_T(H_T))) = 0 \geq F^m(\text{Grim}_T(H_T),\ \bar{s}^m_T(H_T,\ \text{Grim}_T(H_T)))$$

Thus,

$$\text{MaxEPO}^m(T + 1,\ H_{(T+1)},\ \text{Grim}_\infty) = 0.$$

(c) If

$$\text{MinAccept}_T(H_T,\ a^b_T) = \text{NULL},$$

then by the same argument,

$$\text{MaxEPO}^m(T + 1,\ H_{(T+1)},\ \text{Grim}_\infty) = 0.$$

(B) From part (A)(c), above, we can conclude that if

$$\text{MinAccept}_T(H_T,\ a^b_T) = \text{MALICIOUS},$$

then

$$\text{MaxEPO}^m(T,\ H_T,\ \text{Grim}_\infty) > 0,$$

while if

$$\text{MinAccept}_T(H_T,\ a^b_T) = \text{NULL},$$

then

$$\text{MaxEPO}^m(T,\ H_T,\ \text{Grim}_\infty) = 0,$$

and so it must be that choosing NULL cannot be optimal for Mechanical. We are left with two possibilities. Either

$$\text{MaxEPO}^m(T,\ H_T,\ \text{Grim}_\infty) = C + r \times \text{MaxEPO}^m(T + 1,\ H_{(T+1)},\ \text{Grim}_\infty),$$

or

$$\text{MaxEPO}^m(T,\ H_T,\ \text{Grim}_\infty) = D + (1 - p)\, r \times \text{MaxEPO}^m(T + 1,\ H_{(T+1)},\ \text{Grim}_\infty),$$

where

$$H_{(T+1)} \in \mathcal{H}^{\text{coop}}_\infty.$$

But if

$$H_{(T+1)},\ H_{(T+2)} \in \mathcal{H}^{\text{coop}}_\infty,$$

then the period $T + 1$, and $T + 2$ values of the continuation games are identical. Thus, either

$$\text{MaxEPO}^m(T,\ H_T,\ \text{Grim}_\infty) =$$
$$C + r \times (C + r \times \text{MaxEPO}^m(T + 2,\ H_{(T+2)},\ \text{Grim}_\infty)),$$

or

$$\mathrm{MaxEPO}^{m}(T, H_{T}, \mathrm{Grim}_{\infty}) =$$
$$D + (1 - p)r\times(D + (1 - p)r\times\mathrm{MaxEPO}^{m}(T + 2, H_{(T+2)}, \mathrm{Grim}_{\infty})),$$

where

$$H_{(T+1)}, H_{(T+2)} \in \mathcal{H}^{\mathrm{coop}}_{\infty}.$$

Since this also holds in the limit, either

$$\mathrm{MaxEPO}^{m}(T, H_{T}, \mathrm{Grim}_{\infty}) =$$
$$\lim_{\bar{t} \to \infty} \sum_{t=0}^{\bar{t}} [r^{t}\times C + r^{(\bar{t} + 1)}\times\mathrm{MaxEPO}^{m}(\bar{t} + T + 1, H_{(\bar{t} + T+ 1)}, \mathrm{Grim}_{\infty}(H_{\infty}))] =$$
$$\frac{C}{(1 - r)} = \mathrm{EPO}_{C}(\mathrm{Fee}, p),$$

or

$$\mathrm{MaxEPO}^{m}(T, H_{T}, \mathrm{Grim}_{\infty}) =$$
$$\lim_{\bar{t} \to \infty} \sum_{t=0}^{\bar{t}} [(1 - p)^{t} r^{t}\times D + (1 - p)^{(\bar{t} + 1)} r^{(\bar{t} + 1)}\times\mathrm{MaxEPO}^{m}(\bar{t} + T + 1, H_{(\bar{t} + T+ 1)}, \mathrm{Grim}_{\infty}(H_{\infty}))] =$$
$$\frac{D}{(1 - r + rp)} = \mathrm{EPO}_{D}(\mathrm{Fee}, p),$$

which proves the Lemma.

■

The Lemma 6 says that in any period T where the history is cooperative, playing the strategy $\mathrm{MinAccept}_{\infty}$ when the Biological plays Grim_{∞} gives the maximal expected payoff to the Mechanical.

Lemma 6:

$$\forall\ T \in \mathcal{T} \text{ and } \forall\ H_{T} \in \mathcal{H}^{\mathrm{coop}},$$

if

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

then

$$\forall\ \bar{S}^{m}_{\infty} \in \mathcal{S}^{m}_{\infty},$$

it holds that

$$\mathrm{EPO}^{m}(T, H_{T}, B^{b}_{\infty}, S^{m}_{\infty}) \geq \mathrm{EPO}^{m}(T, H_{T}, B^{b}_{\infty}, \bar{S}^{m}_{\infty})$$
$$\mathrm{EPO}^{m}(0, H_{0}, B^{b}_{\infty}, B^{m}_{\infty}) \geq \mathrm{MaxEPO}^{m}(0, H_{0}, B^{b}_{\infty}).$$

Proof:

(A) Suppose for some $T \in \mathcal{T}$,

$$\mathrm{Grim}_{T}(H_{T}) = a^{b}_{T} = (\mathrm{Fee}, p) \in [0, \bar{F}]\times[0, 1]$$

such that

$$\mathrm{EPO}_{C}(\mathrm{Fee}, p) \geq \mathrm{EPO}_{D}(\mathrm{Fee}, p).$$

Then

$$\mathrm{MinAccept}_T(H_T, (\mathrm{Fee}, p)) = a^m_T = \mathrm{CORRECT},$$

and since $H_{(T+1)} \in \mathcal{H}^{coop}$, we get the same outcome in this and all future periods, resulting in a periodic payoff to the Mechanical of $C = \mathrm{Fee} - CP$. This implies that:

$$\mathrm{EPO}^m(T, H_T, \mathrm{Grim}^b_\infty, \mathrm{MinAccept}_\infty) = \mathrm{EPO}_C(\mathrm{Fee}, p) \geq \mathrm{EPO}_D(\mathrm{Fee}, p).$$

Then by Lemma 5,

$$\mathrm{EPO}^m(T, H_T, \mathrm{Grim}^b_\infty, \mathrm{MinAccept}_\infty) = \mathrm{MaxEPO}^m(T, H_T, \mathrm{Grim}^b_\infty).$$

(B) Note that it will never be the case that

$$\mathrm{Grim}_T(H_T) = a^b_T = (\mathrm{Fee}, p) \in [0, \bar{F}] \times [0, 1]$$

such that

$$\mathrm{EPO}_C(\mathrm{Fee}, p) < \mathrm{EPO}_D(\mathrm{Fee}, p).$$

If there is no solution to the Biological's minimization problem, then

$$\mathrm{Grim}_T(H_T) = a^b_T = \mathrm{PASS},$$

and

$$\mathrm{MinAccept}_T(H_T, \mathrm{PASS}) = \mathrm{NULL},$$

and since there is no alternative open to the Mechanical besides to choosing NULL, and

$$\forall\ t > T,\ H_t \notin \mathcal{H}^{coop}_t,$$

it is trivially the case that,

$$\mathrm{EPO}^m(T, H_T, \mathrm{Grim}_\infty, \mathrm{MinAccept}_\infty) = 0 = \mathrm{MaxEPO}^m(T, H_T, \mathrm{Grim}_\infty).$$

(D) We conclude

$$\forall\ \bar{S}^m_\infty \in \mathcal{S}^m_\infty$$

$$\forall\ T \in \mathcal{T} \text{ and } \forall\ H_T \in \mathcal{H}^{coop}_\infty,$$

if

$$B^b_\infty = \mathrm{Grim}_\infty \text{ and } S^m_\infty = \mathrm{MinAccept}_\infty,$$

then

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

and since this also holds for $T = 0$, $H_0 \in \mathcal{H}^{coop}$, it is immediate that:

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

■

The Lemma 7 says that if $S_\infty = (\mathrm{Grim}_\infty, \mathrm{MinAccept}_\infty)$, and this strategy profile is the basis of the belief profile of agents, then (B^b_∞, B^b_∞) satisfies consistency.

Lemma 7:

$$(\mathrm{Grim}_\infty,\ \mathrm{MinAccept}_\infty) = (B^b_\infty,\ B^m_\infty) \in \mathcal{C}^*\mathcal{S}^b_\infty \times \mathcal{C}^*\mathcal{S}^m_\infty.$$

Proof:

(A) First consider Grim_∞.

$$\forall\ T \in \mathcal{T} \ \text{ and } \forall\ H_T \notin \mathcal{H}^{coop}_T$$

it holds that

$$\mathrm{Grim}_T(H_T) = \mathrm{PASS},$$

and

$$\forall\ T \in \mathcal{T} \text{ and } \forall\ H_T \in \mathcal{H}^{coop}_T$$

it holds that

$$\mathrm{Grim}_T(H_T) = (\mathrm{Fee}, p), \ \text{ or } \ \mathrm{PASS},$$

depending on the existence or non-existence, respectively, of a solution an identical minimization problem.

(B) Next consider $\mathrm{MinAccept}_\infty$.

$$\forall\ T \in \mathcal{T},\ \forall\ H_T \notin \mathcal{H}^{coop}_T \ \text{ and } \ a^b_T \in \mathcal{A}^b$$

it holds that

$$\mathrm{MinAccept}_T(H_T,\ a^b_T) = \mathrm{MALICIOUS},\ \text{ or NULL},$$

depending on whether $a^b_T = (\mathrm{Fee}, p) \in [0, \bar{F}] \times [0,1]$, or $a^b_T = \mathrm{PASS}$, respectively, and,

$$\forall\ T \in \mathcal{T} \ \text{ and } \forall\ H_T \in \mathcal{H}^{coop}_T$$

if

$$a^b_T = \mathrm{PASS},$$

then

$$\mathrm{MinAccept}_T(H_T,\ a^b_T) = \mathrm{NULL},$$

while if

$$a^b_T = (\ \mathrm{Fee},\ p) \in [0,\ \bar{F}] \times [0,1],$$

then

$$\mathrm{MinAccept}_T(H_T,\ a^b_T) = \mathrm{CORRECT},\ \text{ or MALICIOUS},$$

depending on the whether $\mathrm{EPO}_C(\mathrm{Fee},\ p)$, or $\mathrm{EPO}_D(\mathrm{Fee}, p)$, respectively, is larger.

Thus, if

$$(\mathrm{Grim}_\infty,\ \mathrm{MinAccept}_\infty) = (B^b_\infty,\ B^b_\infty),$$

then both agents believe that their counterparties will behave identically in essentially identical situations in all periods.

■

Theorem 2 says that the Lemmas proved above imply that the strategy profile

$$S_\infty = (\text{Grim}_\infty, \text{MinAccept}_\infty)$$

is a Consistent Subgame Perfect Equilibrium.

Theorem 2: *If*

$$S^b_\infty = \text{Grim}_\infty \text{ and } S^m_\infty = \text{MinAccept}_\infty,$$

then

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

is a Consistent Subgame Perfect Equilibrium.

Proof:

By Lemma 7,

$$(\text{Grim}_\infty, \text{MinAccept}_\infty) = (B^b_\infty, B^m_\infty) \in \mathcal{C}^*\mathcal{S}^b_\infty \times \mathcal{C}^*\mathcal{S}^m_\infty.$$

By Lemma 2,

$$\forall\ T \in \mathcal{T} \text{ and } \forall\ H_T \notin \mathcal{H}^{\text{coop}},$$

and by Lemmas 4 and 6,

$$\forall\ T \in \mathcal{T} \text{ and } \forall\ H_T \in \mathcal{H}^{\text{coop}},$$

if

$$B^b_\infty = S^b_\infty = \text{Grim}_\infty \text{ and } B^m_\infty = \text{MinAccept}_\infty,$$

then

$$\forall\ \bar{S}^b_\infty \in \mathcal{S}^b_\infty \text{ and } \forall\ \bar{S}^m_\infty \in \mathcal{S}^m_\infty$$

it holds that

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

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

and

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

■

11. References

Acemoglu, Daron, and Pascual Restrepo (2018) Artificial intelligence, automation, and work. In The economics of artificial intelligence: An agenda (pp. 197-236). University of Chicago Press.

AlAshery, Mohamed Kareem, Zhehan Yi, Di Shi, Xiao Lu, Chunlei Xu, Zhiwei Wang, and Wei Qiao. (2020). A blockchain-enabled multi-settlement quasi-ideal peer-to-peer trading framework. *IEEE Transactions on Smart Grid* 12, no. 1: 885-896.

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson (2024) Artificial intelligence, firm growth, and product innovation. *Journal of Financial Economics* 151: 103745.

Ball, Ian and Deniz Kattwinkel (2019). *Probabilistic Verification in Mechanism Desig*n. 389-390. 10.1145/3328526.3329657.

Bebeshko, B., V. Malyukov, M. Lakhno, Pavlo Skladannyi, Volodymyr Sokolov, Svitlana Shevchenko, and M. Zhumadilova. (2022). Application of game theory, fuzzy logic and neural networks for assessing risks and forecasting rates of digital currency. *Journal of Theoretical and Applied Information Technology* 100, no. 24: 7390-7404.

Bichler, Martin, Maximilian Fichtl, Stefan Heidekrüger, Nils Kohring, and Paul Sutterer. (2021). Learning equilibria in symmetric auction games using artificial neural networks. *Nature machine intelligence* 3, no. 8: 687-695.

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicolo, and Sergio Pastorello. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review* 110, no. 10 : 3267-3297

Chaliasos, Stefanos, Marcos Antonios Charalambous, Liyi Zhou, Rafaila Galanopoulou, Arthur Gervais, Dimitris Mitropoulos, and Ben Livshits. (2020). Smart contract and defi security: Insights from tool evaluations and practitioner surveys. *arXiv preprint* arXiv:2304.02981 (2023).

Conley, John (2024) "AI Needs Blockchain: Trustless Solutions to Failures in Machine to Colloidal Markets" *MARBLE Conference Proceedings*, (5th International Conference on Mathematical Research for Blockchain Economy), Forthcoming

Gabriel, Iason. Artificial intelligence, values, and alignment. *Minds and machines* 30, no. 3: 411-437.

Glikson, Ella, and Anita Williams Woolley. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 14, no. 2: 627-660.

Hardjono, Thomas, and Ned Smith.(2021). Towards an attestation architecture for blockchain networks. *World Wide Web* 24, no. 5: 1587-1615.

Hua, Weiqi, Jing Jiang, Hongjian Sun, and Jianzhong Wu. (2020). A blockchain based peer-to-peer trading framework integrating energy and carbon markets. *Applied Energy* 279 : 115539.

Lockey, Steven, Nicole Gillespie, Daniel Holm, and Ida Asadi Someh. (2021). A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions.

Oksanen, Atte, Nina Savela, Rita Latikka, and Aki Koivula. (2020). Trust toward robots and artificial intelligence: An experimental approach to human–technology interactions online. *Frontiers in Psychology* 11: 568256.

Perrett, Cedric, and Simon T. Powers. When to (or not to) trust intelligent machines: Insights from an evolutionary game theory analysis of trust in repeated games. (2021).

Schär, Fabian. (2021). Decentralized finance: On blockchain-and smart contract-based financial markets. FRB of St. Louis Review.

Sima, Violeta, Ileana Georgiana Gheorghe, Jonel Subić, and Dumitru Nancu. (2020). Influences of the industry 4.0 revolution on the human capital development and consumer behavior: *A systematic review. Sustainability* 12, no. 10: 4035.

Tan, Burcu, Edward G. Anderson Jr, and Geoffrey G. Parker. (2020). Platform pricing and investment to drive third-party value creation in two-sided networks. *Information Systems Research* 31, no. 1: 217-239.

Trammell, Philip, and Anton Korinek. (2023). Economic growth under transformative AI. No. w31815. National Bureau of Economic Research.

Wang, Qin, Rujia Li, Qi Wang, and Shiping Chen. (2021). Non-fungible token (NFT): Overview, evaluation, opportunities and challenges. *arXiv preprin*t arXiv:2105.07447

Zarifhonarvar, Ali. (2023) Economics of chatgpt: A labor market view on the occupational impact of artificial intelligence. *Available at SSRN* 4350925

Zeng, Rongfei, Chao Zeng, Xingwei Wang, Bo Li, and Xiaowen Chu. (2021). A comprehensive survey of incentive mechanism for federated learning. *arXiv preprint* arXiv:2106.15406

Zhang, Lejun, Jinlong Wang, Weizheng Wang, Zilong Jin, Yansen Su, and Huiling Chen. (2022). Smart contract vulnerability detection combined with multi-objective detection. *Computer Networks* 217: 109289.

Zhou, Yiyi. (2017). Bayesian estimation of a dynamic model of two-sided markets: Application to the U.S. video game industry. *Management Science* 63, 3874–3894.