Consensus, Distillation, and Trust

On the Mathematics of Agreement in Machines, Networks, and People

Research Draft

Incorporating frameworks from Grunspan & Pérez-Marco (2017),
Hinton et al. (2015), Bandura (1977), and the OMXUS protocol

February 2026

Abstract

This paper explores three manifestations of the same underlying problem: how do distributed agents—whether miners, neurons, or people—arrive at shared agreement without central authority? We begin with Bitcoin's Proof-of-Work consensus, establishing rigorous security bounds via Poisson processes and the regularized incomplete beta function. We then develop a formal analogy between knowledge distillation in neural networks and human behavioural learning, arguing that this connection has underappreciated implications for criminology's rational actor model—not as a dissolution of responsibility, but as a richer account of how behavioural patterns propagate. Finally, we present a scaled-back analysis of social consensus for identity verification and vote counting, treating the OMXUS protocol as a case study in graph-based Sybil resistance. Throughout, we are explicit about where the mathematics is rigorous, where it is merely suggestive, and where open problems remain.

Keywords: blockchain, knowledge distillation, social learning theory, Sybil resistance, Poisson processes, rational actor, criminology

Contents

Part I
Computational Consensus: Bitcoin

1 Introduction

Bitcoin, introduced by Nakamoto in 2008 [1], solved the problem of decentralised monetary consensus: how can a network of mutually distrusting nodes agree on the state of a shared ledger? The answer rests on cryptographic hash functions, Poisson processes, and game-theoretic incentives.

This part presents the standard mathematical treatment. The results are well-established; we include them both for completeness and because they serve as a rigorous benchmark against which the less formal arguments in later parts can be measured.

2 Cryptographic Primitives

Definition 2.1 (Cryptographic Hash Function)**.** *A function $H : \{0,1\}^* \to \{0,1\}^{256}$ is a cryptographic hash function if it satisfies:*

1. ***Preimage resistance**: Given y, it is computationally infeasible to find x with $H(x) = y$.*
2. ***Collision resistance**: It is computationally infeasible to find $x \neq x'$ with $H(x) = H(x')$.*
3. ***Pseudo-randomness**: Changing a single bit of x causes each output bit of $H(x)$ to flip independently with probability $1/2$.*

Bitcoin uses SHA-256 for block hashing, RIPEMD-160 for address derivation, and ECDSA (over the secp256k1 curve) for transaction signatures. The blockchain structure links blocks via hash pointers:

$$\text{BlockHeader} = \big(H(\text{PrevBlock}),\ \text{MerkleRoot},\ \text{Timestamp},\ \text{Difficulty},\ \text{Nonce}\big) \tag{1}$$

where the Merkle root commits to all n transactions with $O(\log n)$ membership proofs.

3 The Mining Process as a Poisson Model

3.1 Exponential Inter-Block Times

Let a miner control fraction $p \in (0,1]$ of total network hashrate, with blocks validated network-wide at average rate one per $\tau_0 = 10$ minutes.

Theorem 3.1 (Mining Time Distribution)**.** *The time T between successive blocks discovered by our miner follows an exponential distribution with rate $\alpha = p/\tau_0$:*

$$f_T(t) = \alpha e^{-\alpha t}, \quad t \geq 0. \tag{2}$$

Proof. The pseudo-random property of H ensures each hash attempt is independent. Let $P(t)$ denote the probability of not finding a valid block by time t. Independence across disjoint intervals gives $P(t+s) = P(t) \cdot P(s)$. The unique continuous solution is $P(t) = e^{-\alpha t}$ for some $\alpha > 0$. Since $\mathbb{E}[T] = \tau_0/p$, we have $\alpha = p/\tau_0$. □

The sum $S_n = T_1 + \cdots + T_n$ of n i.i.d. inter-block times follows a Gamma(n, α) distribution, and the block count $N(t)$ up to time t is Poisson with parameter αt:

$$\mathbb{P}[N(t) = n] = \frac{(\alpha t)^n}{n!}\, e^{-\alpha t}. \tag{3}$$

3.2 The Double-Spend Problem

An attacker with hashrate fraction q (honest fraction $p = 1 - q$) attempts to replace a confirmed transaction by secretly mining an alternative chain.

Theorem 3.2 (Double-Spend Probability; Grunspan & Pérez-Marco, 2017 [2])**.** *After z confirmations, the probability of a successful double-spend by an attacker with $q < 1/2$ is:*

$$P(z) = I_{4pq}(z,\, 1/2) \tag{4}$$

where $I_x(a, b)$ is the regularised incomplete beta function:

$$I_x(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} \int_0^x t^{a-1}(1-t)^{b-1}\,dt. \tag{5}$$

Corollary 3.3 (Exponential Decay)**.** *Setting $s = 4pq < 1$, the double-spend probability decays as:*

$$P(z) \sim \frac{s^z}{\sqrt{\pi(1-s)\,z}} \quad \text{as } z \to \infty. \tag{6}$$

Table 1: Double-spend probability $P(z)$ for various attacker hashrates q.

z	$q = 0.10$	$q = 0.20$	$q = 0.30$	$q = 0.40$
1	0.2045	0.3781	0.5310	0.6696
2	0.0509	0.1531	0.2937	0.4560
3	0.0134	0.0641	0.1653	0.3138
4	0.0036	0.0272	0.0938	0.2175
5	0.0010	0.0117	0.0536	0.1514
6	0.0003	0.0050	0.0308	0.1058

4 Protocol Stability

4.1 Honest Mining is Optimal

Theorem 4.1 (Grunspan & Pérez-Marco, 2018 [3])**.** *In the absence of difficulty adjustment, the revenue-maximising strategy is to publish all blocks immediately upon discovery.*

Proof. Define the revenue ratio $\Gamma = \mathbb{E}[R]/\mathbb{E}[\tau]$ for reward R per cycle of duration τ. Since $N(t)$ is Poisson with intensity $\alpha = p/\tau_0$, the process $M(t) = N(t) - \alpha t$ is a martingale. By Doob's optional stopping theorem, $\mathbb{E}[N(\tau)] = \alpha\,\mathbb{E}[\tau]$ for any stopping time τ, giving $\Gamma \le pb/\tau_0 = \Gamma_H$ with equality only for honest mining. □

4.2 Selfish Mining

The selfish mining attack [4] withholds blocks for strategic advantage. With network connectivity $\gamma \in (0, 1)$, this becomes profitable when $q > (1-\gamma)/(3-2\gamma)$. For average connectivity ($\gamma = 0.5$), the threshold is $q > 0.25$.

Remark 4.2. *This is a genuine vulnerability: the honest-is-optimal result in Theorem 4.1 assumes no difficulty adjustment. Real Bitcoin adjusts difficulty every 2016 blocks, which selfish mining can exploit. The gap between the theoretical and practical security models remains an active research area.*

Part II
Distillation as a Theory of Learned Behaviour

5 Motivation: From Neural Networks to Human Learning

In machine learning, *knowledge distillation* [5] trains a simple "student" network to reproduce the outputs of a complex "teacher." The student never sees the teacher's internal reasoning—only its behavioural outputs (soft probability distributions over outcomes). Yet the student learns to approximate the teacher's competence remarkably well.

We argue this is not merely a metaphor for human behavioural learning—it is structurally the same process. A child learning emotional regulation from a caregiver does not receive explicit rules. Instead, they observe the caregiver's *behavioural outputs* (reactions to situations) and, through repeated exposure and feedback, learn to approximate those responses. The "knowledge" is never explicitly transmitted; it is *distilled* from observed behaviour.

This section formalises the analogy, identifies where it holds rigorously and where it is suggestive, and draws out implications for criminology.

6 Knowledge Distillation: Formal Framework

6.1 The Standard Setup

Definition 6.1 (Knowledge Distillation)**.** *Let $f_T : \mathcal{X} \to \Delta^{C-1}$ be a teacher model mapping inputs to probability distributions over C classes, and $f_S : \mathcal{X} \to \Delta^{C-1}$ a student model. Distillation minimises:*

$$\mathcal{L}_{KD} = (1-\lambda)\,\mathcal{L}_{CE}(f_S, y) + \lambda\, T^2 \cdot \mathrm{KL}\big(\sigma(z_T/T) \parallel \sigma(z_S/T)\big) \tag{7}$$

where z_T, z_S are teacher and student logits, σ is the softmax function, $T > 1$ is a temperature parameter, y is the true label, and $\lambda \in [0,1]$ balances the two objectives.

The temperature T controls how much "dark knowledge" is transferred. At $T = 1$, the student only sees the teacher's hard predictions. As T increases, softer distributions reveal the teacher's uncertainty structure—which wrong answers are "almost right," which classes the teacher finds confusable.

Proposition 6.2 (Dark Knowledge Transfer)**.** *In the high-temperature limit $T \to \infty$, the KL term in* (7) *reduces to matching the teacher's logit differences:*

$$\mathrm{KL}\big(\sigma(z_T/T) \parallel \sigma(z_S/T)\big) \;\approx\; \frac{1}{2CT^2} \sum_{c=1}^{C} (z_{T,c} - z_{S,c})^2 + O(T^{-3}). \tag{8}$$

The student thus learns not just what *the teacher predicts, but the relative structure of the teacher's uncertainty.*

6.2 Graph-Structured Distillation

When the data has graph structure—as in social networks—Graph Neural Networks (GNNs) aggregate information from node neighbourhoods. The Distill n' Explain (DnX) framework [13] distils a GNN teacher into a simpler student, extracting explanations of which neighbourhood structures drive predictions.

Definition 6.3 (Neighbourhood Aggregation)**.** *A GNN with L layers computes node representations via:*

$$h_v^{(\ell+1)} = \phi\left(h_v^{(\ell)}, \bigoplus_{u \in \mathcal{N}(v)} \psi\big(h_v^{(\ell)}, h_u^{(\ell)}\big)\right) \tag{9}$$

where $\mathcal{N}(v)$ is the neighbourhood of v, $\bigoplus$ is a permutation-invariant aggregation (e.g., mean, sum), and ϕ, ψ are learnable functions.

This is where the structural analogy to social learning becomes precise: in a GNN, each node's representation is shaped by its neighbourhood's features through iterated message-passing, just as an individual's behavioural repertoire is shaped by their social environment through repeated interaction.

7 The Human Learning Analogy

7.1 Structure of the Mapping

We now make the analogy explicit. This is not a proof—it is a structural correspondence that we argue is productive.

Table 2: Structural correspondence between knowledge distillation and human behavioural learning.

Component	**Distillation**	**Human Learning**
Teacher	Complex model f_T	Caregiver / social environment
Student	Simple model f_S	Developing individual
Input	Data point $x \in \mathcal{X}$	Situation / stimulus
Output	Soft distribution $\sigma(z/T)$	Behavioural response
Dark knowledge	Inter-class uncertainty	Implicit emotional cues
Temperature	T (softness of signal)	Emotional expressiveness
Loss function	$\mathcal{L}_{\text{KD}}$	Reinforcement / consequence
Graph structure	$\mathcal{N}(v)$ in GNN	Social network

7.2 What This Explains That Rational Choice Does Not

The *rational actor model* in criminology [9, 10] posits that criminal behaviour results from utility-maximising deliberation: an individual weighs expected benefits against expected costs and "chooses" crime when the payoff exceeds the risk.

This model has been productive but struggles with several well-documented phenomena:

1. **Impulsive and emotionally-driven offending**: Many crimes—particularly violent ones—show no evidence of cost-benefit calculation.

2. **Intergenerational transmission**: Criminal behaviour runs in families and communities at rates that exceed what rational imitation would predict.

3. **Context sensitivity**: The same individual behaves very differently across social environments, suggesting behaviour is not driven by stable preferences.

The distillation analogy offers a complementary explanation. If individuals learn behavioural responses by approximating the soft outputs of their social environment—not the explicit rules, but the *implicit distribution over responses*—then:

- Impulsive behaviour corresponds to a student trained at **high temperature**: the individual has absorbed the full uncertainty structure of their environment, including volatile and extreme responses.
- Intergenerational transmission corresponds to **iterated distillation**: each generation distils its behaviour from the previous one, with compounding approximation errors (cf. the "generation loss" problem in repeated distillation [12]).
- Context sensitivity corresponds to **domain shift**: a model trained on one distribution of inputs may behave unpredictably on another.

7.3 Relationship to Existing Theories

We should be clear: the observations above are not new. Bandura's social learning theory [6] established decades ago that behaviour is learned through observation and reinforcement. Sutherland's differential association theory [7] argued that criminal behaviour is learned through interaction with intimate personal groups. Merton's strain theory [8] connected social structure to behavioural outcomes.

The distillation framing does not replace these theories. What it adds is:

1. **A formal mechanism**: the KL-divergence loss provides a precise mathematical description of "approximating observed behaviour," which could in principle be tested against behavioural data.
2. **The dark knowledge insight**: the most important thing transmitted between teacher and student is not the correct answer but the *structure of uncertainty*—which alternatives are "close." Applied to human learning, this suggests that what children absorb from caregivers is not "anger is the right response" but a whole distribution: "anger is likely, withdrawal is possible, calm is unlikely."
3. **Graph-structured propagation**: social learning is not dyadic. It occurs in networks, with neighbourhood aggregation, exactly as in GNNs.

8 On Responsibility

A natural objection: if behaviour is learned through an unconscious distillation-like process, does anyone bear responsibility for their actions?

We think the answer is straightforwardly *yes*, and that the distillation framework actually clarifies why.

Consider: a neural network trained by distillation is responsible for its outputs in the engineering sense—we evaluate it, deploy it or don't, and retrain it if it fails. The fact that it learned its behaviour from a teacher does not absolve the deployed model of producing correct outputs. The teacher's influence is the *causal history*; the model's current weights are the *present reality*.

The same holds for people. You may have acquired a behavioural pattern from your environment decades ago. Understanding that causal history is valuable—it informs intervention, prevention, and rehabilitation. But the pattern is now yours. It operates through your nervous system, produces your actions, and affects other people. The origin of a problem does not determine its ownership.

Remark 8.1 (Practical implication)**.** *This framing suggests that criminal justice should focus less on punishing "rational choices" (which may not have occurred) and more on* ***retraining****—providing new data, new environments, and new feedback signals that update the individual's learned behavioural distribution. This is, in essence, what evidence-based rehabilitation programmes already do.*

8.1 Formal Analogy to Fine-Tuning

In machine learning, a model with poor behaviour can be corrected through *fine-tuning*: continued training on curated data with an appropriate loss function. The analogy to rehabilitation is direct:

$$\mathcal{L}_{\text{rehab}} = \text{KL}\big(\pi_{\text{prosocial}} \parallel \pi_{\text{current}}\big) \tag{10}$$

where $\pi_{\text{prosocial}}$ is a target behavioural distribution and π_{current} is the individual's present response distribution. Minimising this divergence—through structured environments, modelling of alternative responses, and reinforcement—corresponds to updating the "student's" weights toward better behaviour.

The critical point: *fine-tuning works precisely because the model is responsible for its current weights.* If the model were merely a passive conduit for the teacher's outputs, fine-tuning would be incoherent. Responsibility and learned behaviour are not in tension; they are complementary.

Part III
Social Consensus for Identity and Vote Counting

9 Scope and Honest Limitations

Part I established rigorous, well-cited results about Bitcoin. Part II developed a suggestive but not formally proven analogy. This part occupies a middle ground: we present a mathematical framework for social consensus that is internally consistent but whose security guarantees depend on empirical parameters (the "social cost of fraud") that are difficult to measure precisely.

We focus narrowly on the problem OMXUS was designed to address: **counting unique humans for fair resource distribution**, analogous to vote counting. We do not claim the system achieves security "comparable to Bitcoin"—the threat models are fundamentally different, and honest comparison requires acknowledging this.

10 The Sybil Problem in Identity Systems

Definition 10.1 (Sybil Attack [11])**.** *In an identity system, a Sybil attack is the creation of multiple fake identities by a single entity to gain disproportionate influence.*

The fundamental result of Douceur [11] is that without a trusted central authority, Sybil attacks cannot be prevented in a purely open network. All practical defences therefore impose some cost on identity creation.

Bitcoin imposes computational cost (hashrate). OMXUS imposes social cost (vouching from existing verified humans). Neither is absolute; both create economic deterrence.

11 Web-of-Trust Verification

11.1 Graph Model

Definition 11.1 (Verification Graph)**.** *The verification graph is a directed graph $G = (V, E)$ where V is the set of verified identities and $(u, v) \in E$ if u vouched for v. Each vertex has in-degree $d^-(v) \geq k$ (minimum vouch requirement; $k = 3$ in OMXUS).*

Definition 11.2 (Vouch Cost)**.** *Let c_{vouch} denote the cost of a single fraudulent vouch, incorporating:*

$$c_{vouch} = c_{physical} + \mathbb{P}[detection] \cdot c_{penalty} \tag{11}$$

where $c_{physical}$ is the cost of in-person NFC verification and $c_{penalty}$ is the expected loss from detection (trust score reduction, potential ejection).

Proposition 11.3 (Linear Sybil Cost)**.** *Creating n Sybil identities requires at least nk fraudulent vouches. If each vouch has independent cost c_{vouch}, the total attack cost is:*

$$\mathcal{C}(n) \geq n \cdot k \cdot c_{vouch}. \tag{12}$$

Remark 11.4. *This is an* argument*, not a theorem in the sense of Part I. The bound depends on c_{vouch} being accurately estimated, on voucher independence (which collusion violates), and on the detection probability being non-trivial. We state it as a proposition to be honest about its epistemic status.*

11.2 Collusion Detection

Independence among vouchers is the key assumption. The system attempts to enforce it through graph analysis:

Definition 11.5 (Collusion Indicator)**.** *For a set of vouchers S supporting candidate v, define:*

$$\mathcal{I}(S) = \frac{|E(S)|}{\binom{|S|}{2}} \tag{13}$$

where $E(S)$ is the edge set of the subgraph induced by S. High values of $\mathcal{I}(S)$ indicate potential collusion.

This is a heuristic, not a guarantee. Sophisticated attackers can maintain low $\mathcal{I}(S)$ by using vouchers who are not directly connected. The system raises the cost of attack but does not eliminate it.

11.3 Fair Distribution

The narrow claim we *can* make precisely:

Theorem 11.6 (Equal Per-Capita Distribution)**.** *If the verification graph correctly represents unique humans (i.e., no Sybil identities exist), then the distribution mechanism:*

$$r_i(t) = \frac{R(t)}{|V(t)|} \quad \textit{for all } i \in V(t) \tag{14}$$

achieves equal per-capita allocation of resource $R(t)$ at time t.

Proof. Immediate from the definition. The content of the theorem is that the mechanism is *correct given its assumption*. The security question is how well the web-of-trust enforces the assumption. □

Remark 11.7. *This is the honest version of the security claim. The system's fairness is exactly as good as its Sybil resistance, which depends on empirical parameters we cannot pin down with the same precision as Bitcoin's hash-rate-based security.*

12 Bitcoin Anchoring

OMXUS can inherit Bitcoin's well-established security for data integrity by periodically committing state roots to the Bitcoin blockchain.

Definition 12.1 (Epoch Commitment)**.** *For epoch e, the system computes a Merkle Mountain Range root R_e over all identity records and publishes it in a Bitcoin transaction. After z Bitcoin confirmations, reverting R_e requires a Bitcoin double-spend, with probability $I_{4pq}(z, 1/2)$ per Theorem 3.2.*

This is the one place where OMXUS genuinely inherits Bitcoin-grade security: not for identity verification itself, but for the *immutability of committed records*.

Part IV
Synthesis

13 Three Modes of Distributed Agreement

The three parts of this paper address the same abstract problem—distributed consensus—in three different substrates:

Table 3: Three consensus mechanisms and their properties.

	Bitcoin	**Distillation**	**Web-of-Trust**
Agents	Miners	Neurons / people	Verified humans
Agreement on	Transaction ordering	Behavioural distribution	Identity uniqueness
Cost of deviation	Wasted hashrate	Prediction error / consequences	Social cost
Formal guarantees	Strong (Thm 3.2)	Bounded (Prop 6.2)	Conditional (Prop 11.3)
Open problems	Selfish mining gap	Causal identification	Measuring c_{vouch}

14 What Connects Them

The deep connection is that all three systems achieve agreement through **costly signalling**:

- In Bitcoin, the signal is a valid proof-of-work, costly because it requires energy and hardware.
- In distillation/human learning, the signal is behavioural consistency with one's environment, costly because deviation incurs consequences (prediction error for networks; social punishment for people).
- In web-of-trust, the signal is a vouch from an existing member, costly because the voucher stakes their own standing.

In each case, the cost of the signal is what prevents the system from being trivially exploited. The systems differ in how precisely we can quantify that cost—and therefore how rigorous the resulting security guarantees are.

15 Open Problems

1. **Formalising social cost**: Can c_{social} be measured empirically, perhaps through mechanism design experiments? Without this, web-of-trust security arguments remain qualitative.

2. **Testing the distillation analogy**: The correspondence in Table 2 generates testable predictions—e.g., that "high-temperature" social environments (emotionally volatile, inconsistent caregiving) should produce broader behavioural variance. Is this consistent with longitudinal data?

3. **Iterated distillation loss**: In machine learning, repeatedly distilling from student to student causes gradual degradation [12]. Does intergenerational behavioural transmission show analogous "generation loss"? If so, this has direct implications for intervention timing.

4. **The selfish mining gap**: Theorem 4.1 assumes no difficulty adjustment. Closing the gap between the idealised and realistic Bitcoin security models remains open.

5. **Graph-based social learning**: Can GNN message-passing architectures be calibrated against actual social network influence data to test whether neighbourhood aggregation is a good model of human behavioural learning?

16 Conclusion

We have presented three perspectives on distributed consensus, at three levels of mathematical rigour:

1. Bitcoin's Proof-of-Work, where exact probabilities can be computed via the incomplete beta function and protocol optimality can be proven via martingale theory.

2. Knowledge distillation as a model of human behavioural learning, where the formal machinery is suggestive and the structural analogy is precise, but causal claims require empirical validation.

3. Social consensus for identity verification, where the mathematical framework is sound but the key security parameter (c_{vouch}) resists precise quantification.

The value of placing these side by side is not to pretend they are equally rigorous—they are not. It is to show that the *same structural problem* (agreement among distrusting agents) recurs across domains, and that tools from one domain can illuminate another.

The distillation analogy, in particular, offers criminology something concrete: a formal framework for "learning behaviour from observation" that goes beyond metaphor to specify mechanisms, make predictions, and connect to a large body of machine learning theory. Whether this framework ultimately proves empirically productive is an open question. But it is, we think, worth asking.

Consensus is not agreement imposed from above.
It is agreement that emerges from cost, structure, and repeated interaction—
whether the agents are machines, neurons, or people.

References

[1] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," 2008. `https://bitcoin.org/bitcoin.pdf`

[2] C. Grunspan and R. Pérez-Marco, "Double spend races," *Int. J. Theoretical and Applied Finance*, vol. 21, no. 8, 2018.

[3] C. Grunspan and R. Pérez-Marco, "On the profitability of selfish mining," arXiv:1805.08281, 2018.

[4] I. Eyal and E. G. Sirer, "Majority is not enough: Bitcoin mining is vulnerable," *Commun. ACM*, vol. 61, pp. 95–102, 2018.

[5] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv:1503.02531, 2015.

[6] A. Bandura, *Social Learning Theory*. Prentice Hall, 1977.

[7] E. H. Sutherland, *Principles of Criminology*, 4th ed. Lippincott, 1947.

[8] R. K. Merton, "Social structure and anomie," *American Sociological Review*, vol. 3, no. 5, pp. 672–682, 1938.

[9] G. S. Becker, "Crime and punishment: An economic approach," *J. Political Economy*, vol. 76, no. 2, pp. 169–217, 1968.

[10] D. B. Cornish and R. V. Clarke, *The Reasoning Criminal: Rational Choice Perspectives on Offending*. Springer-Verlag, 1986.

[11] J. R. Douceur, "The Sybil attack," in *Int. Workshop on Peer-to-Peer Systems*, 2002, pp. 251–260.

[12] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born again neural networks," in *ICML*, 2018.

[13] P. Longa et al., "Distill n' Explain: Explaining graph neural networks using simple surrogates," in *AISTATS*, 2023.

[14] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman, "SybilGuard: Defending against Sybil attacks via social networks," *ACM SIGCOMM*, 2006.