AGENTSPROTOCOL

Ein dezentrales Protokoll zur semantischen Validierung von Wissen

Version 1.2 — April 2026

Fatih Dinc
fatdinhero@gmail.com
Pforzheim, Deutschland

agentsprotocol.org

Abstract

Das Internet leidet unter einem fundamentalen Defizit: Es gibt keinen universellen, dezentralen Mechanismus, der beweisbar macht, ob eine Information wahr, kontextuell konsistent, relevant und ethisch vertretbar ist. Bestehende Konsensprotokolle ordnen Transaktionen, treffen jedoch keine Aussage ueber die semantische Qualitaet ihres Inhalts. AgentsProtocol schliesst diese Luecke. Es definiert eine protokollare Grundlage — bestehend aus einem semantischen Konsistenz-Score, einem mathematisch bewiesenen Nicht-Kollusions-Test auf Basis der Meta-Bell-Theorie und einem zusammengesetzten WiseScore — auf der jeder Teilnehmer als Validator Wissen beisteuern, validieren und dauerhaft in einem gerichteten azyklischen Graphen verankern kann. Das Resultat ist eine global verfuegbare, faelschungssichere Wissensbasis: eine *Single Source of Truth* fuer das KI-Zeitalter.

Contents

1 Das Problem: Die fehlende Wahrheitsschicht des Internets

Das Internet wurde als Netz des Informationsaustauschs gebaut, nicht als Netz der Wahrheitsfindung. Jede Plattform, jedes KI-System und jede Datenbank hat ihre eigene, proprietaere Sicht auf die Wahrheit — und keine dieser Sichten ist von unabhaengigen Dritten verifizierbar.

Bestehende dezentrale Konsensmechanismen loesen das Problem nur partiell. Proof of Work und Proof of Stake einigen sich auf eine Reihenfolge von Transaktionen, treffen jedoch keine verifizierbare Aussage darueber, ob der Inhalt dieser Transaktionen wahr, kontextuell korrekt oder ethisch vertretbar ist. Fuer reinen Zahlungsverkehr genuegt das. Fuer das Informationszeitalter jedoch nicht.

Neuere Standards wie das *Model Context Protocol* (MCP) von Anthropic oder das *Agent2Agent*-Protokoll (A2A) von Google adressieren die technische Interoperabilitaet zwischen KI-Agenten und externen Diensten. Sie loesen das Problem des sicheren Datenzugriffs und der Agenten-Kommunikation, beantworten jedoch nicht die grundlegendere Frage: *Woher weiss ein KI-Agent — oder ein Mensch —, dass das Wissen, auf das er sich stuetzt, tatsaechlich valide und unabhaengig geprueft wurde?*

> **Das Kernproblem in einer Zeile:** MCP und A2A sind der »USB-C« und das »Bluetooth« des KI-Zeitalters — sie verbinden Systeme. AgentsProtocol ist die »Qualitaetsnorm«, die beweist, dass das Uebertragene auch wahr ist.

Die Folgen dieses Defizits sind weitreichend. KI-Modelle halluzinieren, weil sie keine verifizierbare Grundlage fuer ihre Aussagen haben. Misinformation verbreitet sich, weil kein dezentraler Mechanismus existiert, der Falschaussagen dauerhaft und oeffentlich als solche markiert. Unternehmen und Regulierungsbehoerden investieren enorme Ressourcen in manuelle Pruefprozesse, die fundamental unvollstaendig bleiben.

AgentsProtocol loest dieses Problem durch ein dezentrales, kryptographisch gesichertes Protokoll zur semantischen Validierung — eine »Wahrheitsschicht« fuer das Internet.

2 Vision und Mission

Die **Vision** von AgentsProtocol ist eine global verfuegbare, faelschungssichere Wissensbasis, auf die jeder Mensch, jede KI und jedes autonome System zugreifen kann, um die Qualitaet einer Aussage zu verifizieren. Eine dezentrale *Single Source of Truth*, die keiner zentralen Instanz, keinem Unternehmen und keiner Regierung gehoert.

Die **Mission** lautet: Ein offenes Protokoll zu definieren, das jedem Teilnehmer — unabhaengig von Groesse, Nationalitaet oder Ressourcen — ermoeglicht, als Validator Wissen beizusteuern und validiertes Wissen abzurufen, und das dabei mathematisch beweisbar sicherstellt, dass die Validatoren unabhaengig voneinander arbeiten.

3 Theoretisches Fundament: Die drei Saeulen

AgentsProtocol steht auf drei wissenschaftlich ausgearbeiteten Fundamenten, die jeweils als eigenstaendige Whitepaper veroeffentlicht wurden und als integraler Bestandteil dieser Spezifikation gelten.

darkblue	
Dezentralisierung	Kein einzelner Akteur kontrolliert die Wissensbasis.
lightgray Beweisbarkeit	Jede Aussage traegt einen kryptographischen Qualitaetsnachweis.
Unabhaengigkeit	Meta-Bell-Statistik beweist die Unabhaengigkeit der Validatoren.
lightgray Offenheit	Jeder kann Node betreiben, Claims einreichen, Wissen abfragen.
Neutralitaet	Das Protokoll ist wertneutral; Anwendungsdomaenen definieren ihre Ethik-Parameter.
lightgray Haltbarkeit	Validiertes Wissen wird dauerhaft in einem DAG verankert.

Table 1: Kerneigenschaften von AgentsProtocol

3.1 Meta-Bell-Theorie (MBT)

Die Meta-Bell-Theorie ist das mathematische Fundament des gesamten Systems. Sie definiert ein masstheoretisches Verschraenkungsmas Ψ, das quantifiziert, wie stark die beobachteten Korrelationen zwischen mehreren Parteien von jeder moeglichen lokalen Hidden-Variable-Erklaerung abweichen.

Im Kontext von AgentsProtocol bedeutet dies: Wenn mehrere Validatoren voneinander unabhaengig arbeiten, erzeugen sie statistisch unkorrelierte Fehlermuster. Wenn sie hingegen kollaborieren — sei es durch geteilte Modelle, geteilte Daten oder explizite Absprache — erzeugen sie identische Muster. Der Ψ-Wert unterscheidet diese beiden Szenarien mathematisch beweisbar:

$$\Psi(X,Y) = \max_{\lambda \in \Lambda} \frac{\left| E_{\text{beobachtet}}(X,Y) - E_{\text{klassisch}}(X,Y \mid \lambda) \right|}{\Delta_{\text{krit}}}$$

- $\Psi = 0$: Vollstaendige Kollusion oder geteilte versteckte Variable — Block wird abgelehnt.
- $\Psi = 1$: Vollstaendige Unabhaengigkeit — maximaler Vertrauensnachweis.
- $\Psi \geq \Psi_{\min}$: Akzeptanzschwelle, protokollspezifisch festgelegt.

Das klassische CHSH-Ungleichungssystem ist der Spezialfall trivialer Stoerung. Die MBT verallgemeinert dies zu einem universellen Rahmen, der auf jedes System korrelierter Zufallsvariablen angewendet werden kann — also insbesondere auf ein Netzwerk von Validatoren.

3.2 Proof of WiseWork (PoWW)

Proof of WiseWork definiert den Qualitaetsmass stab fuer Informationseinheiten. Eine Informationseinheit ist ein Vier-Tupel $i = (v, c, r, e)$, bestehend aus einem Wahrheitskandidaten, einem Kontextgewicht, einem Relevanzfaktor und einem Ethik-Konformitaetswert.

Der **WiseScore** einer Einheit ist das Produkt aus vier normalisierten Komponenten:

darkblue		
$T(i)$ — **Wahrheit**	$\frac{\exp(\alpha \cdot v_i)}{\sum_j \exp(\alpha \cdot v_j)}$	Normalisierter Wahrheitsscore via Maximum-Entropie-Prinzip.
lightgray $C(i)$ — **Kontext**	$\frac{c_i}{\sum_j c_j}$	Anteil des Kontextgewichts an der Gesamtmenge.
$R(i)$ — **Relevanz**	$\log(1 + r_i)$	Logarithmisch gedaempft; verhindert Relevanzinflation durch Angreifer.
lightgray $E(i)$ — **Ethik**	$e_i \in [0, 1]$	Konformitaet mit der domaenenspezifischen Wertebasis.

Table 2: Komponenten des WiseScore

$$W(i) = T(i) \cdot C(i) \cdot R(i) \cdot E(i)$$

Die **multiplikative Form** ist entscheidend: Eine Einheit muss in allen vier Dimensionen hinreichend gut bewertet sein. Eine ethisch nicht vertretbare, aber technisch wahre Aussage erhaelt einen Score nahe null. Additive Formen wuerden erlauben, schwache Dimensionen durch starke zu kompensieren.

Der aggregierte Block-Score ist das arithmetische Mittel aller WiseScores im Block:

$$\text{PoWW} = \frac{1}{|I|} \sum_{i \in I} W(i)$$

3.2.1 Domaenenspezifische Ethik

Die Ethik-Konformitaet $E(i)$ bezieht sich auf eine **domaenenspezifische Wertebasis**, die im Claim-Kontext deklariert wird. Das AgentsProtocol selbst schreibt keine universellen Werte vor; es validiert lediglich die Uebereinstimmung mit der vom Einreicher gewaehlten Referenz.

3.3 Proof of Independent Semantic Validation (PoISV)

PoISV ist das operative Protokoll, das PoWW und MBT zu einem lauffaehigen Konsensmechanismus verbindet. Es fuehrt den **semantischen Konsistenz-Score** $S_{\mathbf{con}}$ ein, der deterministisch berechnet, wie konsistent ein Claim mit einem oeffentlichen, versionierten Wissenskorpus κ ist.

3.3.1 Praezise Definition von $S_{\mathbf{con}}$

Der S_{con}-Score quantifiziert die Uebereinstimmung eines Claims mit κ in drei Schritten:

1. **Extraktion:** Aus dem Claim-Text werden mittels eines vortrainierten Sprachmodells (z. B. eines Satz-Transformers) die Entitaeten, Relationen und der Aussagekern als Embedding-Vektor $\mathbf{v}_A \in \mathbb{R}^d$ extrahiert.
2. **Abruf:** Aus κ werden alle Fakten abgerufen, die dieselben Entitaeten betreffen. Diese werden ebenfalls als Embedding-Vektoren $\{\mathbf{v}_\kappa^{(1)}, \ldots, \mathbf{v}_\kappa^{(m)}\}$ repraesentiert.

3. **Aehnlichkeitsmass:** Der Score ist definiert als

$$S_{\text{con}}(A) = \max\left(0,\ \frac{\cos\big(\mathbf{v}_A,\ \bar{\mathbf{v}}_\kappa\big) - \tau}{1 - \tau}\right)$$

wobei $\bar{\mathbf{v}}_\kappa = \frac{1}{m}\sum_{j=1}^{m} \mathbf{v}_\kappa^{(j)}$ der mittlere Vektor der abgerufenen Fakten ist, $\cos(\cdot,\cdot)$ die Kosinus-Aehnlichkeit, und $\tau \in [0, 1)$ ein protokollspezifischer Schwellwert.

Die verwendeten Modelle und der Wert von τ sind Teil der Protokollversion und werden als Hash im Blockheader verankert. Dies gewaehrleistet vollstaendige Reproduzierbarkeit und Ueberpruefbarkeit.

3.3.2 Nicht-Kollusions-Test Ψ (operativ)

Jeder Validator loest k kanonische Kontroll-Claims $D_1, \ldots, D_k$ mit bekannten Loesungen $S^*(D_j)$ und erzeugt den Fehlervektor

$$\mathbf{e}_i = \big(|S_i(D_j) - S^*(D_j)|\big)_{j=1}^{k}.$$

Fuer N Validatoren in einem Block ist die **gewichtete** Ψ-Statistik (zur Sybil-Resistenz, siehe Abschnitt 4.3):

$$\Psi = 1 - \frac{\displaystyle\sum_{1 \le i < j \le N} w_i\, w_j\, |\rho(\mathbf{e}_i, \mathbf{e}_j)|}{\displaystyle\sum_{1 \le i < j \le N} w_i\, w_j}$$

mit $w_i = \sqrt{s_i}$, wobei s_i der vom Validator hinterlegte Stake ist. Unabhaengige Validatoren erzeugen unkorrelierte Fehlermuster ($\Psi \approx 1$); kolludierende Validatoren erzeugen identische Muster ($\Psi \approx 0$).

3.3.3 Akzeptanzregel

Ein Block wird genau dann akzeptiert, wenn:

$$\frac{1}{|A|}\sum_{A \in \text{Block}} S_{\text{con}}(A) \ \ge\ \theta_{\min} \quad \text{und} \quad \Psi \ \ge\ \Psi_{\min}.$$

4 Das AgentsProtocol: Architektur und Komponenten

4.1 Rollen im Netzwerk

4.2 Lebenszyklus eines Claims

1. **Einreichung:** Ein Claim $A = (d, \sigma)$ wird mit den serialisierten Aussagedaten und einer optionalen digitalen Signatur an alle Validatoren uebertragen.

2. **Reception-Schicht:** Jeder Validator prueft den Claim gegen lokale Eingangsregeln, schreibt Zeitstempel und Signatur fest.

3. **Comprehension-Schicht:** Jeder Validator berechnet $S_{\text{con}}(A)$ gegen κ sowie die Komponenten T, C, R, E und W fuer die Einheit.

darkblue		
Claim-Einreicher	Reicht eine Aussage (Claim) zur Validierung ein. Kann jede Person, KI oder jedes System sein.	Digitale Signatur
lightgray Validator / Node	Berechnet S_{con} und $W(i)$ fuer eingehende Claims, loest Kontrollaufgaben, schlaegt Bloecke vor.	Node-Software, Stake-Hinterlegung
Leichter Client	Prueft die Blockakzeptanz anhand von Headern und zk-Beweisen.	Download der Blockheader

Table 3: Rollen im AgentsProtocol-Netzwerk

4. **Kontrollaufgaben:** Jeder Validator loest k kanonische Kontroll-Claims und erzeugt seinen Fehlervektor $\mathbf{e}_i$ fuer den Ψ-Test.

5. **Cognition-Proof-Schicht:** Ein Validator, der einen Block vorschlagen moechte, erzeugt einen Zero-Knowledge-Beweis π ueber die Korrektheit der gesamten Berechnung (via zkVM).

6. **Konsens:** Andere Validatoren akzeptieren den Block, wenn die Scores die Schwellenwerte erfuellen und der zk-Beweis verifiziert. Akzeptierte Bloecke werden dem DAG hinzugefuegt.

4.3 Sybil-Resistenz durch gewichtetes Staking

Ein Angreifer koennte versuchen, den Ψ-Wert zu manipulieren, indem er eine grosse Zahl scheinbar unabhaengiger Validator-Identitaeten (Sybils) erstellt, die jedoch alle dasselbe kolludierende Verhalten zeigen. Um dies zu verhindern, koppelt AgentsProtocol die Teilnahme am Konsens an einen **Stake** — eine Hinterlegung von AGENTS-Token, die als Sicherheitspfand dient.

Die Ψ-Statistik wird als gewichtete mittlere absolute Pearson-Korrelation berechnet. Die Quadratwurzel-Gewichtung $w_i = \sqrt{s_i}$ stellt sicher, dass der Einfluss eines Validators sublinear mit seinem Stake waechst — ein Angreifer muesste ueberproportional viel Kapital einsetzen, um den Ψ-Wert signifikant zu beeinflussen. Diese Konstruktion verhindert Sybil-Angriffe ohne Rueckgriff auf zentrale Identitaetspruefungen.

4.4 DAG-Ordnungsschicht

AgentsProtocol nutzt das **GHOSTDAG**-Protokoll als Ordnungsschicht. Dieses erlaubt parallele Blockproduktion ohne den Verlust ehrlicher Arbeit: Jeder neue Block verweist auf alle sichtbaren Spitzen als Elternbloecke, und ein k-Cluster-Algorithmus induziert eine totale Ordnung.

Das Gewicht eines Blocks bestimmt seine Stellung in der kanonischen Kette:

$$\text{Gewicht}(B) = \Psi_B \cdot \sum_{A \in B} S_{\text{con}}(A)$$

Die kanonische Kette ist stets der Pfad durch den DAG mit dem hoechsten kumulierten Gewicht.

4.5 Zero-Knowledge-Beweise (modular)

Die Korrektheit der Score-Berechnungen wird durch eine **generische zkVM** gesichert. Als Referenzimplementierung dient die Nexus zkVM; das Protokoll ist jedoch so spezifiziert, dass **jede RISC-V-basierte zkVM mit succinct proofs** kompatibel ist (z. B. RISC Zero, SP1). Die verwendete zkVM-Version wird im Blockheader festgehalten.

Die rohen Evidenzquellen und Modellgewichte der Validatoren verbleiben als privater Zeuge in der zkVM; nur die finalen Scores und die Ψ-Statistik werden veroeffentlicht. Dies gewaehrleistet einerseits die Verifizierbarkeit der Berechnung und andererseits den Schutz vertraulicher Validator-Daten.

4.6 Privatspheare

Fuer jede Claim-Einreichung sollte ein neues Schluesselpaar verwendet werden. Der Ψ-Test operiert ausschliesslich auf der Korrelationsstruktur der oeffentlichen Validator-Ausgaben und benoetigt keinerlei Inhaltsdaten — er liefert damit einen Nicht-Kollusions-Nachweis, ohne selbst zusaetzliche Informationen preiszugeben.

5 Sicherheitsanalyse

5.1 Angriffsszenarien

Ein rationaler Angreifer mit Kontrolle ueber einen Anteil $q < 0.5$ der Validatoren steht vor einer einfachen Kosten-Nutzen-Rechnung. Er kann seine Ressourcen verwenden, um sich als ehrlicher Validator zu verhalten und regulaere Blockertraege zu erhalten — oder er kann versuchen, eine alternative Kette mit manipulierten Inhalten aufzubauen.

Die zweite Option ist aus zwei Gruenden ausserordentlich schwer. Erstens muss sein Block beide Schranken (Score und Ψ) gleichzeitig ueberschreiten. Zweitens kann er die Ψ-Schranke nicht durch Abstimmung erfuellen, weil koordinierte Validatoren identische Fehlermuster erzeugen und damit die Ψ-Statistik auf null druecken. Er waere gezwungen, *tatsaechlich unabhaengige* Validatoren zu betreiben — was den Zweck der Koordination unterlaeuft.

5.2 Quantitative Sicherheitsgarantien

Die Erfolgswahrscheinlichkeit eines Angriffs faellt exponentiell in zwei unabhaengigen Dimensionen: der Anzahl der Bloecke z, die der Angreifer aufholen muss, und der Groesse der erforderlichen kollusiven Kohorte k.

darkblue			
0.10	32	0.7	$< 10^{-12}$
lightgray 0.20	32	0.7	$< 10^{-8}$
0.30	64	0.7	$< 10^{-7}$
lightgray 0.40	64	0.7	$< 10^{-4}$
0.49	128	0.7	$< 10^{-3}$

Table 4: Erfolgswahrscheinlichkeit eines Angriffs in Abhaengigkeit von q, k und $\Psi_{\min}$

Die kombinierte Sicherheitsgarantie ergibt sich als Produkt der Score-Schranke (Nakamoto-Sicherheit) und der Ψ-Schranke (Meta-Bell-Sicherheit). Waehrend die Score-Schranke bei hohem q an Kraft verliert, wird die Ψ-Schranke genau dann exponentiell staerker, weil k mit q waechst.

Ein Angreifer mit $q = 0.49$ und sechs Bloecken Rueckstand bei hundert Validatoren pro Block erreicht eine gemeinsame Erfolgswahrscheinlichkeit unter 10^{-23}.

5.3 Sicherheit bei Mehrheitsangriffen

Sollte ein Angreifer mehr als 50 % des Gesamtstakes kontrollieren ($q \geq 0.5$), kann er die **Ordnung** der Bloecke manipulieren — er kann legitime Bloecke zensieren oder die Kette reorganisieren. Dies entspricht der bekannten 51 %-Attacke in Proof-of-Work- und Proof-of-Stake-Systemen.

Jedoch kann ein solcher Angreifer **nicht** die semantische Validitaet der einzelnen Claims verfaelschen. Da die S_{con}-Berechnung oeffentlich und deterministisch ist und der Wissenskorpus κ ueber inhaltsadressierbare Hashes referenziert wird, kann jeder leichte Client einen ungueltigen Claim als solchen erkennen. Der Schaden ist somit auf **Zensur** und **temporaere Verwirrung** beschraenkt, nicht auf die dauerhafte Einschleusung falschen Wissens.

Darueber hinaus ist die oekonomische Huerle fuer einen 51 %-Angriff extrem hoch: Der Angreifer muesste mehr als die Haelfte aller im Umlauf befindlichen AGENTS-Token erwerben und staken. Ein solcher Kapitaleinsatz wuerde den Token-Preis massiv in die Hoehe treiben, wodurch der Angriff unwirtschaftlich wird. Zudem wuerde ein erfolgreicher Angriff das Vertrauen in das gesamte Protokoll zerstoeren und den Wert der erbeuteten Token vernichten — ein klassisches *Miner's Dilemma*.

6 Token-Oekonomie und Anreizstruktur

AgentsProtocol verwendet eine native Waehrung, den **AGENTS-Token**, um Validatoren fuer ihre semantische Arbeit zu entlohnen und das Netzwerk in einem dezentralen, langfristig stabilen Gleichgewicht zu halten. Die maximale Gesamtmenge ist unwiderruflich auf **1.000.000.000 (eine Milliarde) AGENTS** festgelegt. Diese harte Obergrenze schafft nachweisbare Verknappung und verhindert eine unkontrollierte Inflation – ein entscheidender Faktor fuer die langfristige oekonomische Sicherheit des Netzwerks.

6.1 Belohnungsmechanismus

Der erste Eintrag eines jeden Blocks ist eine spezielle Coinbase-Transaktion, die neu gepragte AGENTS-Token an den Validator auszahlt, der den Block erfolgreich vorgeschlagen hat. Die Belohnung pro Block ist proportional zum gewichteten Score: Je hoeher der durchschnittliche S_{con}-Score der enthaltenen Claims und je naeher Ψ bei 1 liegt, desto groesser faellt die Ausschuettung aus.

Zusaetzlich koennen Claim-Einreicher eine **Gebuehr** (in AGENTS) entrichten, um ihre Claims priorisiert validieren zu lassen. Diese Gebuehren fliessen vollstaendig an den vorschlagenden Validator.

6.1.1 Halbierungsplan (Halving)

Die Ausgabe neuer Token folgt einem deterministischen, abnehmenden Zeitplan, der Verknappung und Planbarkeit kombiniert:

- **Initiale Blockbelohnung:** 100 AGENTS pro Block.
- **Halbierung:** Alle 2.100.000 Bloecke (ca. 4 Jahre bei einer Blockzeit von 60 Sekunden) wird die Belohnung halbiert.

- **Langfristige Anreize:** Sobald der Grossteil der 400 Mio. Validator-Token ausgegeben ist, uebernehmen Transaktionsgebuehren die Rolle als primaere Einnahmequelle der Validatoren.

6.2 Anreizkompatibilitaet

Das System ist anreizkompatibel konstruiert: Ein rationaler Akteur findet keine Moeglichkeit, durch Manipulation der Scores oder des Ψ-Werts mehr zu verdienen als durch ehrliche, unabhaengige Validierungsarbeit. Ein Angriff gefaehrdet nicht nur den eigenen Stake, sondern wuerde den Wert aller AGENTS-Token zerstoeren — ein klassisches *Miner's Dilemma*.

6.3 Token-Distribution

Die Verteilung der 1.000.000.000 AGENTS-Token auf die verschiedenen Akteursgruppen ist wie folgt festgelegt. Die fuer Validator-Belohnungen reservierten Token werden gemaess dem Halbierungsplan ueber Jahrzehnte ausgegeben; alle uebrigen Kategorien werden beim Mainnet-Launch generiert und unterliegen den angegebenen Sperrfristen.

darkblue		
Validator-Belohnungen (Halbierungsplan)	40 %	400.000.000 AGENTS
lightgray Oekosystem-Fonds (Grants, Entwicklung)	25 %	250.000.000 AGENTS
Gruenderteam & fruehe Beitragende*	15 %	150.000.000 AGENTS
lightgray Community-Airdrop (Bootstrapping)	10 %	100.000.000 AGENTS
Reserve (Upgrades, Notfaelle)	10 %	100.000.000 AGENTS
Gesamt	**100 %**	**1.000.000.000 AGENTS**

* Vesting-Periode: 4 Jahre mit 1 Jahr Cliff.

Table 5: AGENTS-Token-Distribution (Gesamtangebot: 1 Milliarde)

7 Governance und Protokoll-Entwicklung

AgentsProtocol ist als offenes, dezentrales Protokoll konzipiert, das keiner einzelnen Instanz gehoert.

7.1 On-Chain-Governance

Aenderungen am Protokoll — insbesondere an den Akzeptanzschwellen $\theta_{\min}$ und $\Psi_{\min}$ sowie an der Token-Oekonomie — werden durch Token-Abstimmung entschieden. Grosse Token-Holder haben mehr Stimmgewicht, doch die MBT-Ψ-Statistik wird auch auf Abstimmungsmuster angewendet, um koordinierte Stimmkauf-Angriffe zu erkennen.

7.1.1 Governance der Kontrollaufgaben

Das Set $D_1, \ldots, D_k$ der kanonischen Kontroll-Claims wird durch einen on-chain Abstimmungsprozess der Token-Inhaber festgelegt. Die Referenzloesungen $S^*(D_j)$ sind als Hash im Genesisblock verankert. Neue Aufgaben koennen per Proposal eingebracht werden und erfordern eine Karenzzeit von mindestens vier Wochen, um *Teaching to the Test* zu verhindern.

7.2 Off-Chain-Governance

Die technische Weiterentwicklung des Protokolls erfolgt durch einen offenen Spezifikationsprozess aehnlich dem Bitcoin Improvement Proposal (BIP)-System. Jede Protokollaenderung muss durch ein veroeffentlichtes Dokument mit formaler Begruendung und Sicherheitsanalyse untersetzt sein.

7.3 Institutionelle Struktur

Es wird empfohlen, AgentsProtocol unter einer gemeinnuetzigen Stiftung (vorzugsweise in der Schweiz oder Deutschland) zu verankern. Die Stiftung haelt das Protokoll-IP, foerdert die Open-Source-Entwicklung und dient als neutraler Ansprechpartner fuer Regulierungsbehoerden. Die Stiftung hat kein Vetorecht ueber Protokollaenderungen, die durch die Community entschieden werden.

8 Roadmap

Hinweis: Die angegebenen Zeitraeume sind Ziele und koennen sich aufgrund technischer, regulatorischer oder gemeinschaftlicher Entwicklungen verschieben. Die Roadmap wird kontinuierlich im oeffentlichen GitHub-Repository aktualisiert.

9 Abgrenzung zu bestehenden Protokollen

AgentsProtocol konkurriert nicht mit bestehenden Protokollen — es ergaenzt sie auf einer fundamentaleren Ebene.

Ein KI-Agent, der ueber MCP auf eine Datenbank zugreift, kann die erhaltenen Informationen gegen AgentsProtocol pruefen, um zu verifizieren, ob sie von unabhaengigen Validatoren geprueft wurden. Die Protokolle arbeiten in unterschiedlichen Schichten und sind aufeinander aufbaubar.

10 Anwendungsbeispiele

10.1 Verifizierbare KI-Antworten

Ein KI-Assistent kann jede seiner Antworten mit einem AgentsProtocol-Beweis versehen: *»Diese Aussage wurde von einem dezentralen Netzwerk mit einem WiseScore von 0.97 und einem Unabhaengigkeitsnachweis* $\Psi = 0.89$ *validiert.«* Nutzer koennen diesen Beweis selbst verifizieren.

10.2 Dezentrale Faktenpruefdatenbank

Medienorganisationen, Regierungsbehoerden und zivilgesellschaftliche Akteure koennen Claims in das Netzwerk einreichen. Das Ergebnis ist eine oeffentlich zugaengliche, faelschungssichere Datenbank validierter Fakten — ohne einen zentralen Gatekeeper.

darkblue		
Phase 0: Fundament	Q2 2026	Veroeffentlichung Whitepaper, GitHub-Repository, technische Spezifikation, Domain agentsprotocol.org.
lightgray Phase 1: Spezifikation	Q3 2026	Claim-Schema (JSON), API-Endpunkte, Kontrollaufgaben-Set v1, Pseudocode-Implementierung, Community-Aufbau.
Phase 2: Proof of Concept	Q4 2026 – Q1 2027	Rudimentaerer Validator-Client (Rust/Go), Simulation des Ψ-Tests, erste externe Entwickler-Beitraege, Grant-Antraege.
lightgray Phase 3: Testnet	Q2 – Q4 2027	Oeffentliches Testnet, zkVM-Integration, Sicherheitsaudit, Token-Modell-Simulation.
Phase 4: Mainnet	Q1 – Q2 2028	Mainnet-Launch, Token-Distribution, erste kommerzielle Integrationen, Governance-Aktivierung.
lightgray Phase 5: Oekosystem	2028+	SDK fuer Entwickler, API-Gateway, Integration in KI-Frameworks, regulatorische Anerkennung (EU AI Act).

Table 6: Entwicklungs-Roadmap von AgentsProtocol

10.3 EU AI Act Compliance

Der EU AI Act verlangt von Hochrisiko-KI-Systemen, dass ihre Outputs nachvollziehbar und geprueft sind. AgentsProtocol liefert einen kryptographisch belegten Audit-Trail fuer jede Ausgabe eines KI-Systems — und erleichtert damit die Konformitaetspruefung erheblich.

10.4 Halal-Compliance im Finanzsektor

Die Ethik-Komponente $E(i)$ des WiseScore kann domaenenspezifisch konfiguriert werden. Fuer islamische Finanzprodukte koennte eine spezialisierte Ethik-Wertebasis (AAOIFI-Standards) als Validierungsgrundlage dienen — und jede Aussage ueber die Halal-Konformitaet eines Finanzprodukts dezentral und beweisbar validiert werden.

10.5 Wissenschaftliche Publikationen

Forschungsergebnisse koennen als Claims eingereicht und von unabhaengigen Experten-Validatoren geprueft werden. Das Ergebnis ist ein dezentrales Peer-Review-System, das schneller, transparenter und resistenter gegen korporative Einflussnahme ist als bestehende Verfahren.

darkblue		
Bitcoin / PoW	Konsens ueber Transaktionsreihenfolge	Inhaltliche Wahrheitsfindung
lightgray Ethereum / PoS	Konsens + Smart Contracts	Semantische Validierung
MCP (Anthropic)	Interoperabilitaet KI ↔ Dienste	Verifizierbarkeit des Inhalts
lightgray A2A (Google)	Agenten-Kommunikation	Wahrheitsnachweis
AgentsProtocol	**Semantische Validierung + Wahrheitsnachweis + Kollusions-Beweis**	**Alle o. g. Faelle sind komplementaer**

Table 7: Abgrenzung von AgentsProtocol zu bestehenden Protokollen

11 Schlussfolgerung

Das Internet braucht eine Wahrheitsschicht. Nicht im Sinne einer zentralen Instanz, die vorschreibt, was wahr ist — sondern im Sinne eines offenen, dezentralen Protokolls, das jedem Teilnehmer die Werkzeuge gibt, die Qualitaet einer Aussage selbst zu pruefen.

AgentsProtocol definiert diese Schicht. Auf dem mathematischen Fundament der Meta-Bell-Theorie, dem Qualitaetsmass stab des Proof of WiseWork und dem operativen Rahmen des Proof of Independent Semantic Validation entsteht eine globale, faelschungssichere Wissensbasis — kontrolliert von niemandem, zugaenglich fuer jeden.

Dies ist keine ferne Vision. Die theoretischen Grundlagen sind ausgearbeitet. Die Spezifikation liegt vor. Die Domain `agentsprotocol.org` ist gesichert. Der naechste Schritt ist die Gemeinschaft: Entwickler, Forscher und Institutionen, die erkennen, dass die Wahrheitsfindung des 21. Jahrhunderts dezentral, beweisbar und offen sein muss.

Mitmachen

agentsprotocol.org | GitHub: `agentsprotocol/specification` | Kontakt: `fatdinhero@gmail.com`

References

[1] F. Dinc, *Meta-Bell-Theorie: Eine masstheoretische Erweiterung der Bell-Ungleichungen. Grundlagen, Dynamik und statistische Inferenz.* SHA-256: 062e290009f6b7339e9a8b522ce1d94d9021d109d8e4bc41210d1f3dda053a3b, 2026.

[2] F. Dinc, *Proof of WiseWork: Ein Peer-to-Peer-System zur Konsensbildung ueber Wahrheit.* Version 2.0, SHA-256: e57cae993701a1933a3317e28c7bb7141a01b51b31674adee97b6cf89472c2eb, 2026.

[3] F. Dinc, *PoISV: Ein Peer-to-Peer-System zur unabhaengigen semantischen Validierung.* Version 1.0, April 2026.

[4] S. Nakamoto, *Bitcoin: A Peer-to-Peer Electronic Cash System.* 2008.

[5] J. S. Bell, *On the Einstein Podolsky Rosen paradox.* Physics 1(3), 195–200, 1964.

[6] J. F. Clauser, M. A. Horne, A. Shimony, R. A. Holt, *Proposed experiment to test local hidden-variable theories.* Physical Review Letters 23, 880–884, 1969.

[7] Y. Sompolinsky, S. Wyborski, A. Zohar, *PHANTOM GHOSTDAG: A scalable generalization of Nakamoto consensus.* AFT 2021.

[8] Nexus Labs, *Nexus zkVM: Enabling Verifiable Computation.* https://nexus.xyz, 2024.

[9] C. E. Shannon, *A mathematical theory of communication.* Bell System Technical Journal 27, 379–423, 1948.

[10] W. Feller, *An introduction to probability theory and its applications.* John Wiley & Sons, 1957.