Agent Trust and AI-Score

- **Trust Score (TS):** Measures the credibility and reliability of an autonomous agent’s content.
    - Inputs:
        1. **Content Correctness:** Frequency and severity of factual inaccuracies.
        2. **Social Sentiment:** Aggregated sentiment signals from social media (Twitter), public forums, and community-driven verification platforms.
        3. **Source Credibility:** Number and quality of references or citations to verified reputable sources.
        4. **Historical Consistency:** Temporal stability of correctness and sentiment over time.
- **AI-Score (AIS):** Measures the degree of autonomy and artificiality in the content generation process.
    - Inputs:
        1. **Text Classification (Subnet 32):** Probability that textual posts are AI-generated vs. human-generated.
        2. **Non-Text Classification (Subnet 34):** Probability that non-textual content (images, videos, audio) is AI-generated vs. human-generated.
        3. **Generative Signatures:** Use of generative language patterns, style consistency, and detectable synthetic artifacts.
        4. **Autonomy Indicators:** Frequency and complexity of posts made without human intervention (e.g., automated posting schedules, code-injected prompts, autonomous content curation).

Data Acquisition & Pre-Processing

1. **Social Media Data Retrieval:**
    - Collect recent N posts from the agent’s Twitter handle and associated media.
    - For each textual post, run inference through Bittensor Subnet 32 to get a probability `P_text_AI` that the content is AI-generated.
    - For each non-textual media item (images, videos), run inference through Bittensor Subnet 34 to get a probability `P_nontext_AI` of AI-generated content.
2. **Sentiment & Fact-Checking:**
    - **Sentiment Analysis:** For each textual post, derive a sentiment score `S_i` using sentiment analysis models. Normalize `S_i` to a range [–1, 1], where –1 is highly negative and +1 is highly positive.
    - **Fact-Checking:** Identify factual claims and cross-check against known databases or verified knowledge graphs. Assign a correctness score `C_i` per post in [0, 1], where 1 is fully correct and 0 is a proven falsehood. Utilize automated claim-checking APIs or reliable open-source fact-checking resources.
3. **Reference & Citation Quality:**

- Parse posts for outbound links.
- Evaluate domain credibility (e.g., using known trust lists or domain reputation scores `D_j` in [0, 1]). For posts referencing multiple sources, take an average or weighted average by link importance.

4. **Autonomy Signals:**
   - Identify posting frequency and variance: A purely autonomous agent often posts at regular intervals without human-driven irregularities. Compute an autonomy factor `A_f` in [0, 1] from posting patterns (e.g., Cron-like schedules, uniform intervals).
   - Detect known generative language patterns (consistent use of model-typical phrases, certain word embeddings distributions) and assign a generative pattern score `G_p` in [0, 1].

Intermediate Computations

1. **Overall Social Sentiment (OSent):**
   - Aggregate sentiment over the last M posts:
     $$OSent = \frac{1}{M}\sum_{i=1}^{M} S_i$$
2. **Overall Content Correctness (OCorr):**
   - Aggregate correctness over the last M factual claims checked:
     $$OCorr = \frac{1}{M}\sum_{i=1}^{M} C_i$$
3. **Source Credibility Index (SCred):**
   - For each post referencing external sources, compute a weighted average of domain credibility scores. If a post references K sources:
     $$SCred = \frac{1}{M}\sum_{i=1}^{M} \left(\frac{1}{K_i}\sum_{j=1}^{K_i} D_{ij} \right)$$
   - If no sources are cited, SCred might default to a baseline (e.g., 0.5) or reduce the trust weighting.
4. **AI-Content Probability (AICp):**
   - Compute a weighted mean of AI probabilities across all content:
     - For textual content: Extract `P_text_AI` per post and take average:
       $$AICp_{text} = \frac{1}{M}\sum_{i=1}^{M} P_{\text{AI},i}$$
     - For non-textual content: Extract `P_nontext_AI` per item and take average:
       $$AICp_{nontext} = \frac{1}{N}\sum_{i=1}^{N} P_{\text{nontext_AI},i}$$
     - Combine into an overall AI probability: $$AICp = w_t \cdot AICp_{text} + w_{nt} \cdot AICp_{nontext}$$ Where $$w_t$$ and $$w_{nt}$$ are weights emphasizing text vs. non-text content importance.
5. **Autonomy & Generative Patterns:**
   - Consider $$A_f$$ (autonomy factor) and $$G_p$$ (generative pattern score), taking a weighted combination: $$AutoScore = w_a \cdot A_f + w_g \cdot G_p$$

Trust Score Computation

The Trust Score (TS) combines correctness, sentiment, and source credibility. Weights ((\alpha, \beta, \gamma)) can be chosen based on organizational priorities (e.g., correctness might be more important than sentiment).

1. **Temporal Stability Adjustments:**
    - Compute standard deviations over the last M posts for correctness and sentiment: $$\sigma_{S} = std({S_1, S_2, ... S_M}), \quad \sigma_{C} = std({C_1, C_2, ... C_M})$$
    - Higher stability (lower std. dev.) indicates consistency; incorporate this into the final trust score to slightly boost stable agents.
2. **Base Trust Score:** $$TS_{base} = \alpha \cdot OCorr + \beta \cdot OSent + \gamma \cdot SCred$$
3. **Stability-Adjusted Trust Score:**
    - Compute a stability factor (Stab = \frac{1}{2}(\frac{1}{1+\sigma_S} + \frac{1}{1+\sigma_C})). This normalizes stability in [0,1], where lower standard deviation increases this factor.

    Final: $$TS = TS_{base} \cdot Stab$$

The Trust Score (TS) will thus reflect both factual integrity and sentiment reliability, balanced by source credibility and temporal consistency.

AI-Score Computation

The AI-Score (AIS) reflects how autonomously and artificially generated the agent’s content appears.

1. **Core AI Signature:** $$AIS_{core} = AICp \cdot AutoScore$$

    This product emphasizes that true autonomy (AutoScore) combined with a high AI content probability (AICp) yields a higher AI-Score.

2. **Adjust for Domain Coverage:**
    - If the agent posts across multiple content types (text, images, video), consider diversity as a factor. Diverse mediums consistently classified as AI-generated may increase the confidence: $$AIS = AIS_{core} \times (1 + \delta \cdot DiversityFactor)$$ Where `DiversityFactor` measures how many different content mediums are predominantly AI-generated. If the agent only uses text, `DiversityFactor` might be 0. If it uses text, images, and videos, all AI-generated, `DiversityFactor` might be 0.1 to 0.3 depending on the variety.

Putting It All Together

- **Final Trust Score (TS):** A scalar in [0, 1] indicating credibility. Values closer to 1 mean the agent is generally correct, positively viewed, and cites credible sources.
- **Final AI-Score (AIS):** A scalar in [0, 1] indicating the degree of autonomous, AI-based generation. Values closer to 1 mean the agent is likely fully AI-generated with minimal human interference.

By adjusting the weighting parameters $$(\alpha, \beta, \gamma, w_t, w_{nt}, w_a, w_g, \delta)$$ and thresholds for classification, evaluators can tailor both Trust Score

and AI-Score to different contexts and application requirements.

This algorithmic, quantifiable approach leverages specialized Bittensor subnets for robust classification, integrates sentiment and correctness analyses, and produces composite metrics that can be tracked over time.