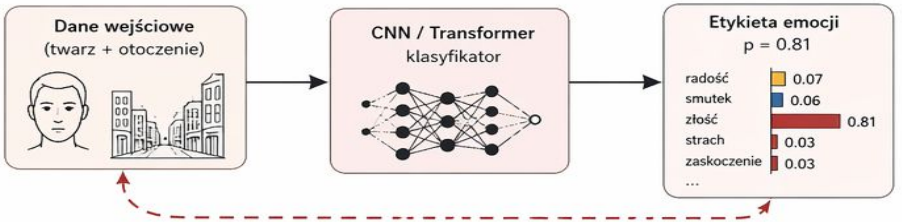
Architektury emocji

1 PRZEPŁYW INFORMACJI I PODEJMOWANIE DECYZJI

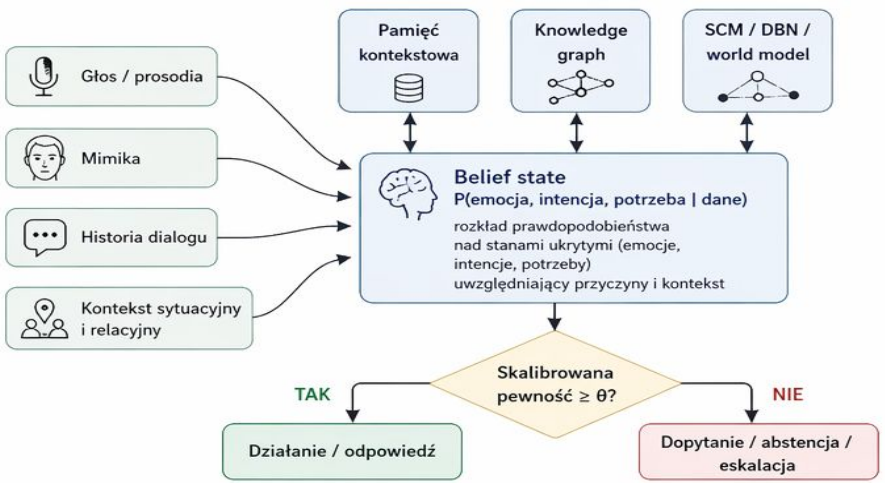
✕ SYSTEM KORELACYJNY

Podejście korelacyjne (czarna skrzynka)



✓ ARCHITEKTURA PRZYCZYNOWO-KONTEKSTOWA

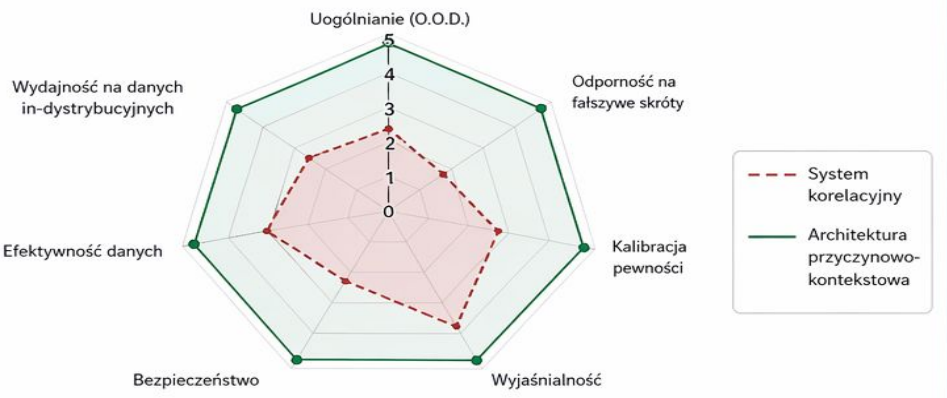
Wnioskowanie z uwzględnieniem przyczyn i kontekstu



2 PORÓWNANIE WŁAŚCIWOŚCI ARCHITEKTUR

KRYTERIUM	SYSTEM KORELACYJNY (czarna skrzynka)	ARCHITEKTURA PRZYCZYNOWO-KONTEKSTOWA
Źródło sygnału	Płaskie cechy z danych (statystyczne współwystępowanie)	Wielomodalne + kontekst + wiedza (przyczyny)
Mechanizm	Dopasowanie wzorca (fitting korelacji)	Wnioskowanie przyczynowe (model świata / SCM / DBN)
Uogólnianie (O.O.D.)	Słabe	Silne (lepsza odporność na zmianę kontekstu)
Odporność na fałszywe skróty	Niska	Wysoka
Kalibracja pewności	Często źle skalibrowana	Jawna, możliwa do kalibracji
Wyjaśnialność	Niska (nieprzejrzyste decyzje)	Wysoka (jawne przyczyny i zależności)
Bezpieczeństwo	Niskie (brak mechanizmu abstencji)	Wysokie (abstencja / eskalacja / weryfikacja)

PROFIL ARCHITEKTUR (skala 0–5)



Affective loop jako rdzeń interakcji

1. REPREZENTACJE EMOCJI – JAK MODELE ROZUMIEJĄ AFEKT?

A. MODELE DYSKRETNE

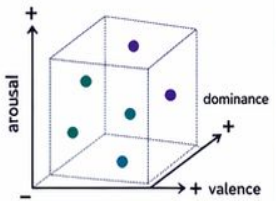
Kategorie podstawowe
(np. Ekman, Plutchik)



+ interpretowalność
– sztywne granice

B. MODELE WYMIAROWE

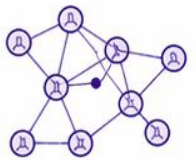
Przestrzeń ciągła
(np. valence–arousal–dominance)



+ uchwyt dynamiki
– trudniejsza interpretacja

C. MODELE KONSTRUKTYWISTYCZNE

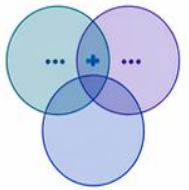
Emocje jako konstrukty
znaczeniowe i społeczne



+ kontekst i znaczenie
– zależność od ramy

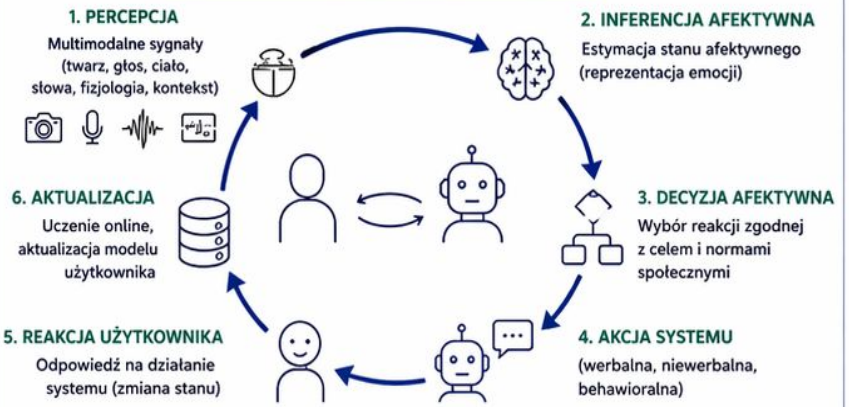
D. MODELE HYBRYDOWE

Połączenie kategorii,
wymiarów i kontekstu



+ elastyczność
i bogactwo
– złożoność obliczeniowa

2. AFFECTIVE LOOP – RDZEŃ INTERAKCJI



Ciągłe współ-kształtowanie stanu afektywnego i relacji

3A. ŚRODOWISKO TESTOWE (LABORATORYJNE)

- ✓ Dobre oczyszczenie danych
- ✓ Czytelne wyrażenia emocji
- ✓ Ograniczona pula uczestników
- ✓ Często neurotypowi uczestnicy
- ✓ Uproszczone warunki sensoryczne
- ✓ Kontrolowany kontekst zadania

METRYKI KLASYCZNE



Sensowne w warunkach i.i.d.
i przy stabilnej dystrybucji danych

PRZEJŚCIE DO RZECZYWISTEGO WDROŻENIA

Distribution Shift
i złożoność świata

3B. ŚRODOWISKO WDROŻENIA (REAL-WORLD)

- ⚠ Afekt miesza się z zadaniem i kontekstem
- ⚠ Przeciążenie sensoryczne, zakłócenia
- ⚠ Historia interakcji i oczekiwania
- ⚠ Różnice indywidualne i kulturowe
- ⚠ Szum, ambiguity i niepełne obserwacje
- ⚠ Distribution shift (czasu, miejsca, populacji)

WCZORAJ TRAFNY MODEL MOŻE:

- ⊗ generować nieadekwatne reakcje
- ⊗ być nierobustny wobec nowych użytkowników
- ⊗ utrwalać uprzedzenia i wykluczenia
- ⊗ nie wychwycić realnego distresu
- ⊗ nadmiernie interweniować lub intrudować

METRYKI NIEWYSTARCZAJĄCE



Nie odzwierciedlają skutków
społecznych i bezpieczeństwa

4. PORAŻKA W ROBOTYCE AFEKTYWNEJ NIE JEST NEUTRALNA

FAŁSZYWY DISTRESS (FP)



Ryzyko: intruzywna interwencja,
utrata zaufania, eskalacja stresu

FAŁSZYWE ZAANGAŻOWANIE (FN)



Ryzyko: przeoczenie przeciążenia,
pogorszenie dobrostanu, możliwa szkoda

PYTANIE METODOLOGICZNE, NIE TYLKO TECHNICZNE:

„Czy model jest trafny?"

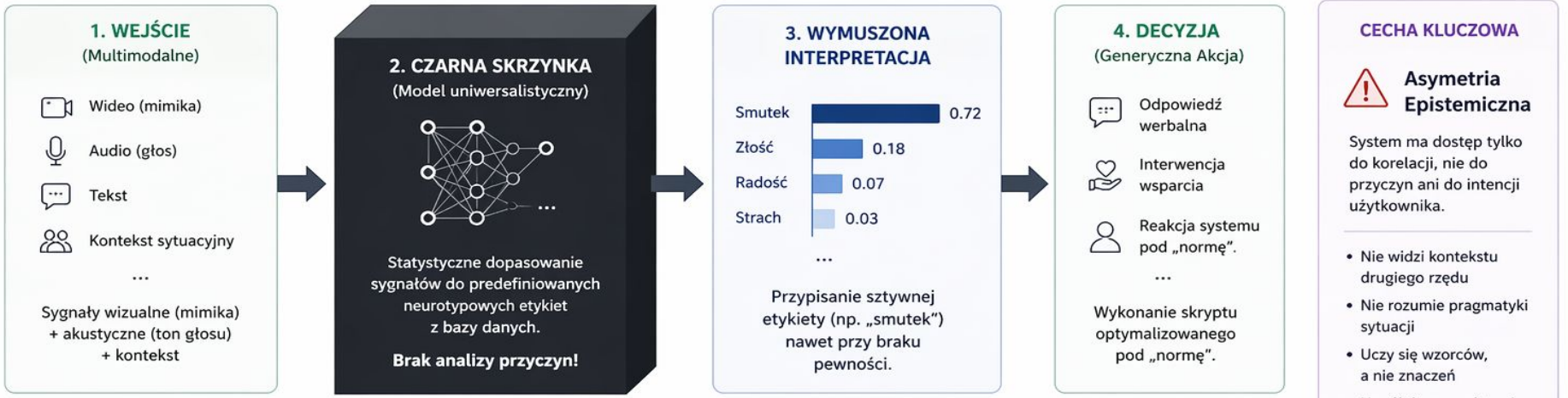
czy raczej

„Czy potrafi fail gracefully
w sytuacji społecznie wrażliwej?"

5. ZASADY PROJEKTOWE DLA PĘTLI AFEKTYWNEJ

- **Uprzywilejowanie bezpieczeństwa społecznego** – minimalizacja ryzyk niekorzystnych skutków.
- **Kalibracja niepewności i świadomość ograniczeń** – pewność modelu ≈ pewność decyzji.
- **Projektowanie dla abstencji i dopytania** – „nie wiem" jest lepsze niż błąd kosztowny.
- **Personalizacja i uczenie relacyjne** – poznawanie użytkownika w czasie.
- **Wielowymiarowa ewaluacja** – skuteczność, dobrostan, zaufanie, sprawiedliwość, nierówności.

Asymetria epistemiczna w „Empatii przez korelację”



CECHA KLUCZOWA

Asymetria Epistemiczna

System ma dostęp tylko do korelacji, nie do przyczyn ani do intencji użytkownika.

- Nie widzi kontekstu drugiego rzędu
- Nie rozumie pragmatyki sytuacji
- Uczy się wzorców, a nie znaczeń
- Uogólnia na podstawie statystycznej średniej, nie wyjątków

NA CZYM POLEGA ASYMETRIA EPISTEMICZNA?

SYSTEM (Ograniczone epistemicznie)

Dane obserwowalne (dystrybucje) → Korelacje statystyczne $p(x|y)$ → Sztywne etykiety / klasy (etyikiety neurotypowe)

Brak dostępu do stanów ukrytych, intencji, znaczeń i kontekstu pragmatycznego!

UŻYTKOWNIK (Bogaty kontekst i intencje)

Stany wewnętrzne (emocje, potrzeby, intencje, cele) → Kontekst (pragmatyka, relacje, historia, kultura) → Znaczenie sytuacji (intencja komunikacyjna, znaczenie drugiego rzędu)

KONSEKWENCJE

- Błędne etykietowanie
- Niewłaściwe interwencje
- Utrata zaufania
- Nadmierne / niedostateczne działania
- Utrwalanie uprzedzeń i biasów

JAK ZMNIEJSZYĆ ASYMETRIĘ?

Modelowanie przyczynowe
SCM / DBN / world model aby odróżnić korelacje od przyczyn.

Aktywne dopytywanie
Zadawaj pytania, aby zmniejszyć niepewność epistemiczną.

Użytkownik w pętli
Współtworzenie znaczeń, walidacja interpretacji, opcje użytkownika.

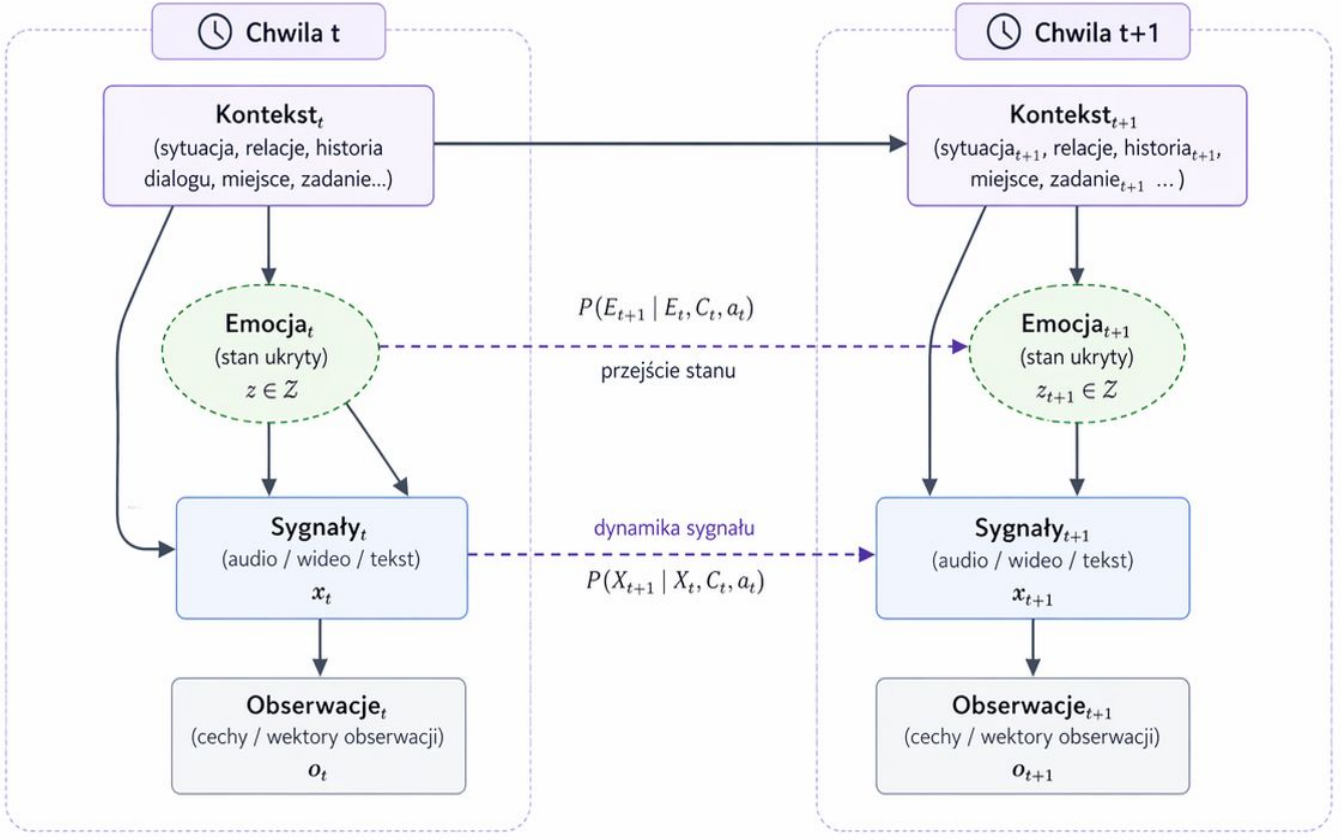
Kalibracja i abstencja
Znaj własne ograniczenia. Abstencja, gdy niepewność jest wysoka.

Dostosowanie do kontekstu
Uwzględniaj normy, kulturę, relacje i historię interakcji (pełny kontekst).

PRZESŁANIE

Empatia przez korelację ≠ zrozumienie. Prawdziwa empatia wymaga zmniejszania asymetrii epistemicznej.

Dynamiczny model emocji, dynamic Bayesian Network



Legenda

- **Kontekst** — zmienna obserwowalna
- **Emocja (stan ukryty)** — zmienna ukryta
- **Sygnały** — zmienna obserwowalna (wielomodalna)
- **Obserwacje** — zmienna obserwowalna (cechy)
- Zależność w obrębie tej samej chwili
- Zależność czasowa / przejście probabilistyczne

Notacja

- z_t – stan emocjonalny (ukryty)
- C_t – kontekst w chwili t
- x_t – wektor sygnałów (wielomodalny)
- o_t – wektor obserwacji (cechy)
- a_t – działanie agenta (opcjonalnie kontrolujące przejścia i sygnały)

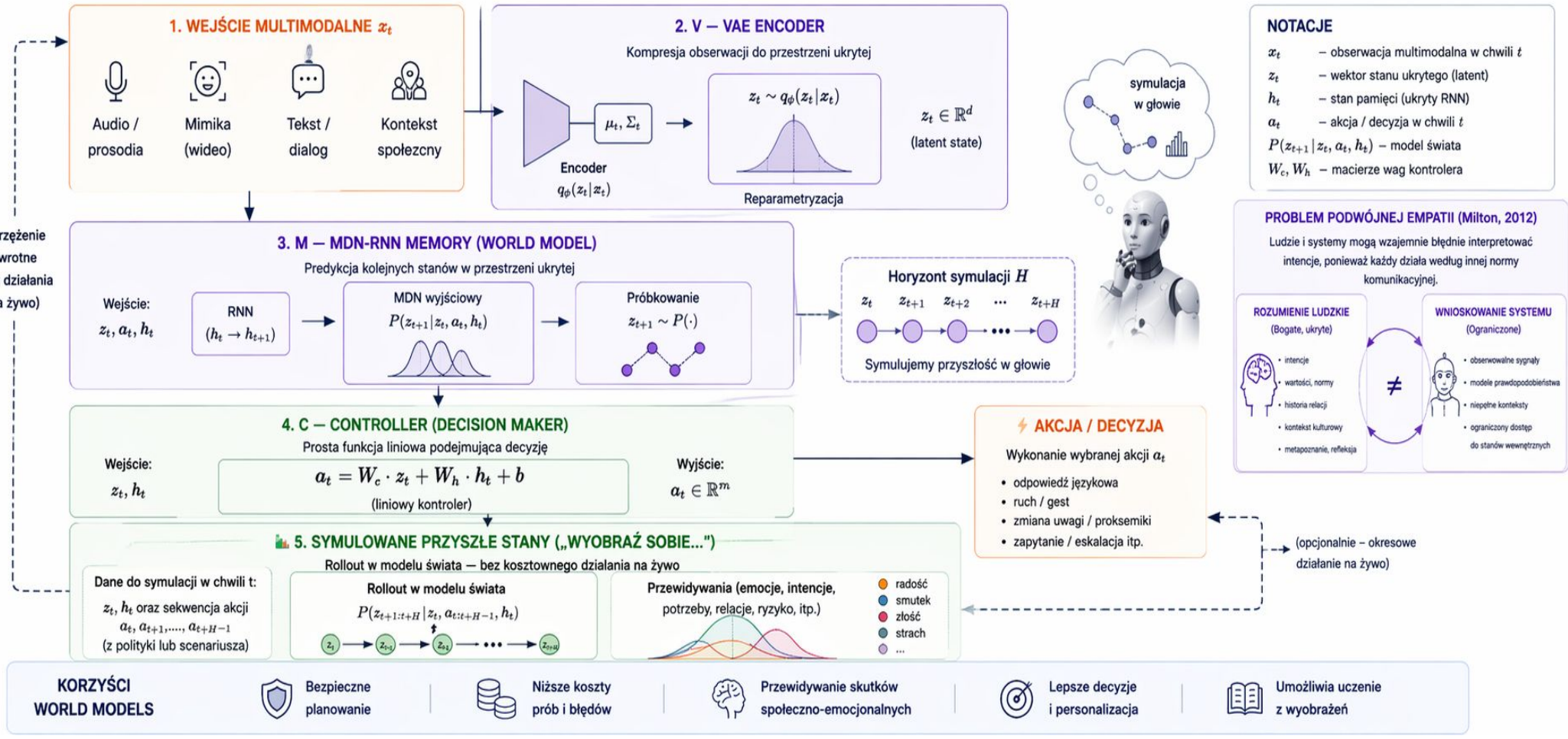
Faktoryzacja wspólnej dystrybucji (łańcuch Markowa 1. rzędu)

$$P(z_{1:T}, x_{1:T}, o_{1:T} | C_{1:T}, a_{1:T-1}) = P(z_1 \mid C_1) \prod_{t=1}^{T-1} P(z_{t+1} \mid z_t, C_t, a_t) \prod_{t=1}^{T} P(x_t \mid z_t, C_t, a_t) P(o_t \mid x_t)$$

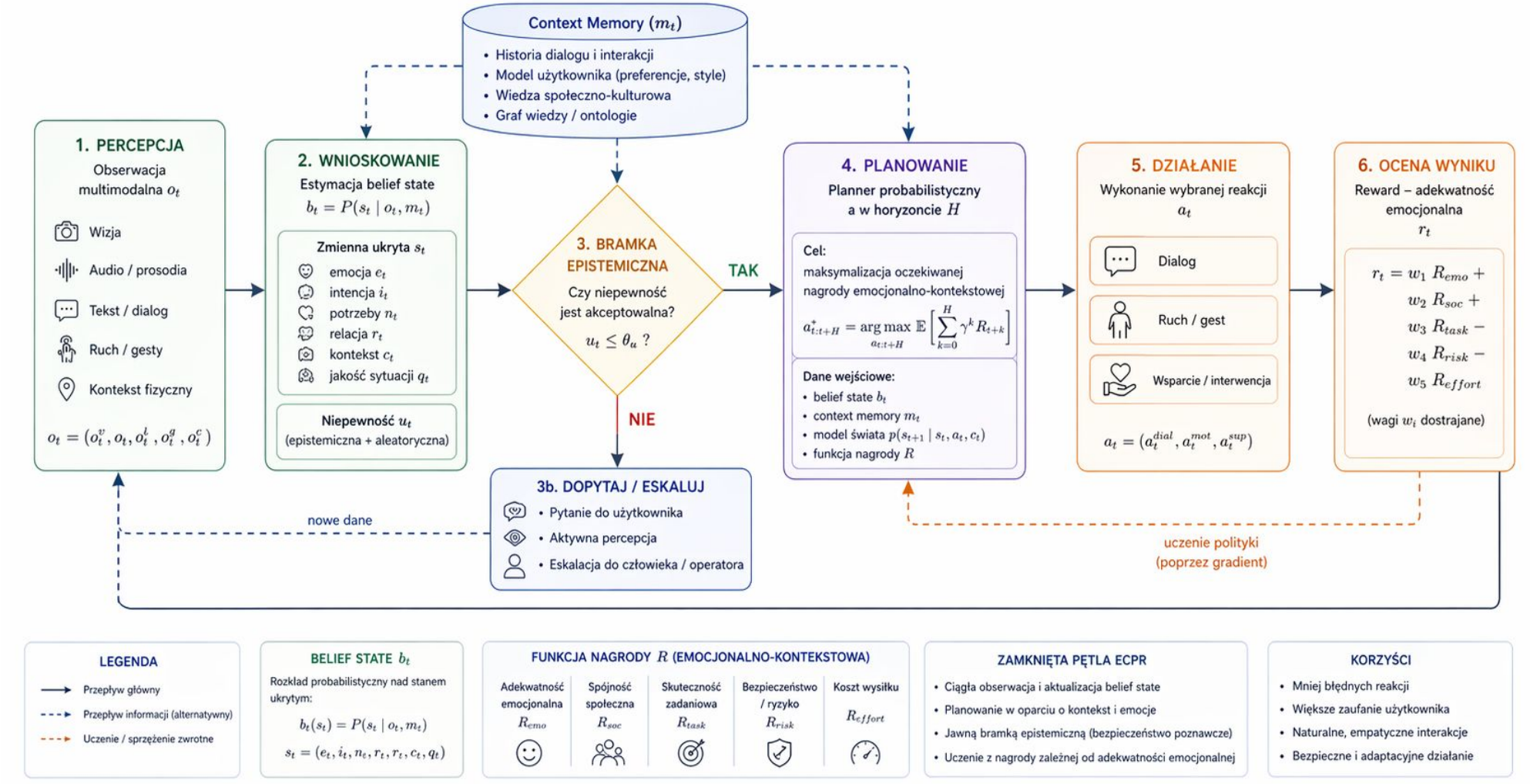
Założenia modelu

- Emocja jest stanem ukrytym zmieniającym się w czasie.
- Kontekst wpływa zarówno na emocję, jak i na generację sygnałów.
- Sygnały (wielomodalne) są generowane przez emocję w danym kontekście.
- Obserwacje są funkcją sygnałów (np. cech ekstrakcyjnych).

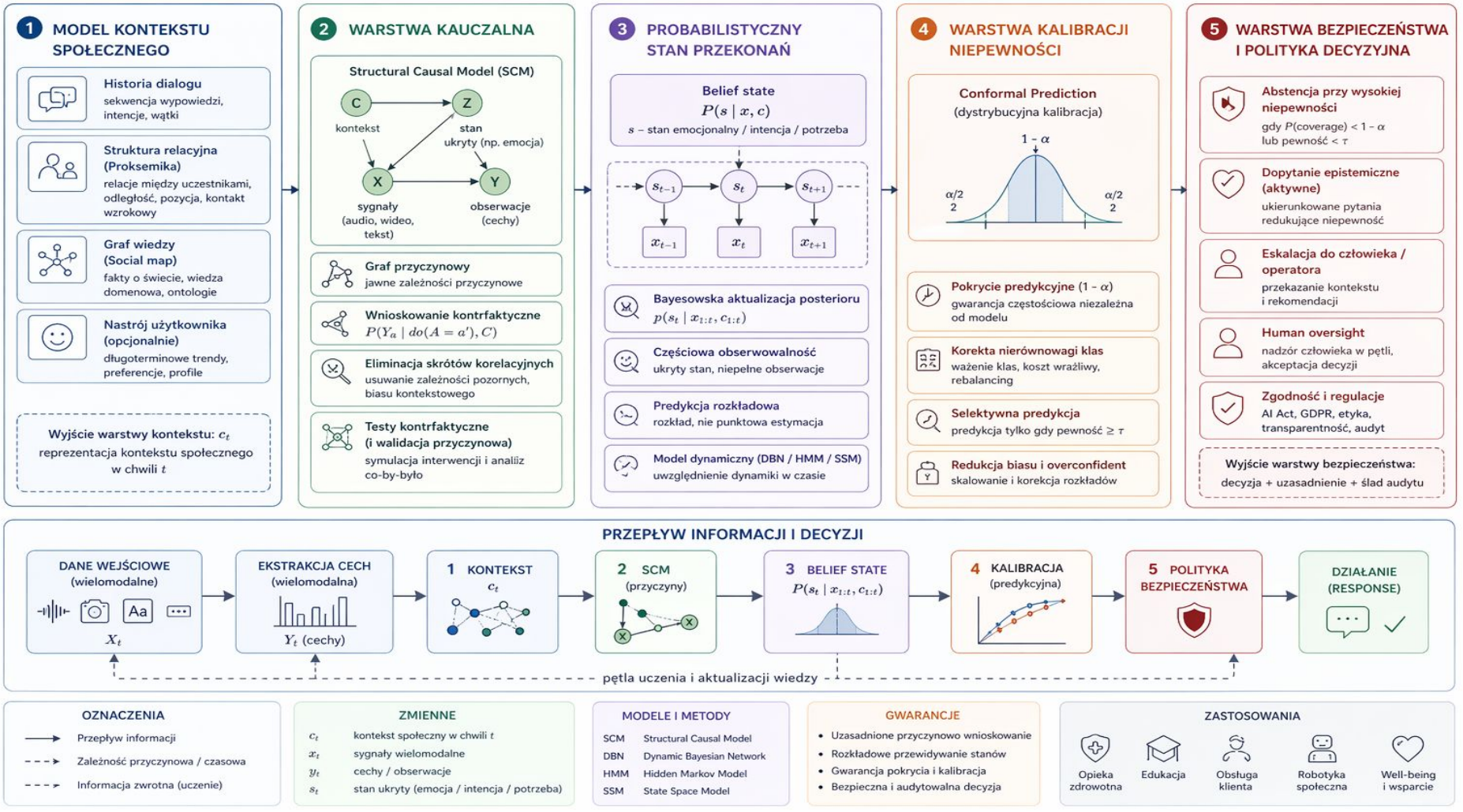
World models: VAE →MDM-RNN → Controller



ECPR: Emotion-context reinforced planner

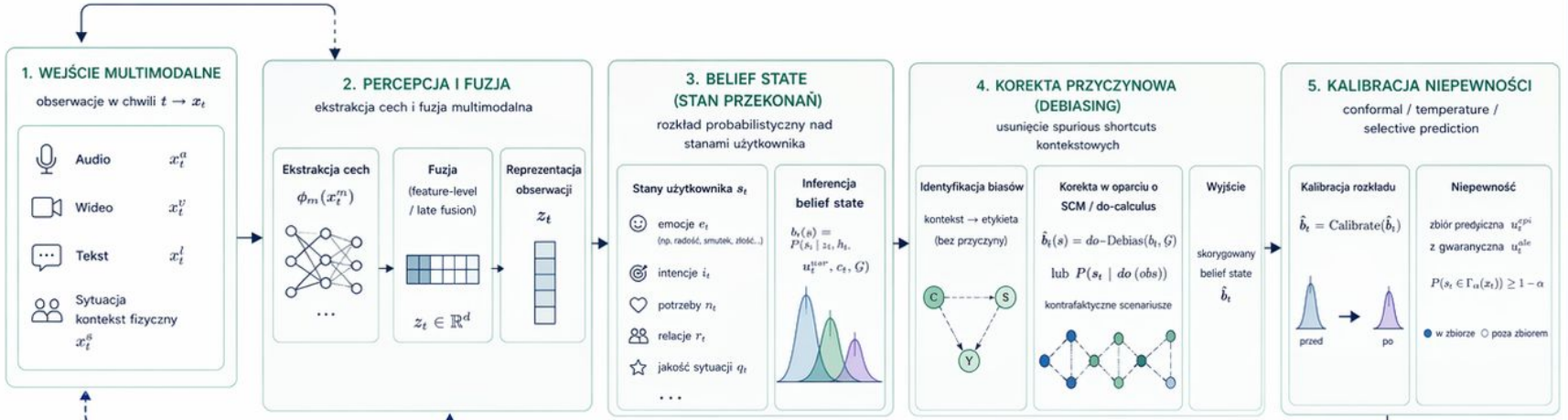


CASIE-2: Architektura przyczynowo-kontekstowa emocji



Architektura z epistemiczną pokorą

Pełny potok decyzyjny: od wejścia multimodalnego do działania, dopytania lub abstencji



A. PAMIĘĆ DIALOGU I MODEL UŻYTKOWNIKA

- Historia dialogu $h_t = (u_{1:t-1}, a_{1:t-1})$
- Model użytkownika u_t^{usr} (preferencje, profil, rola, zaufanie)
- Stan relacji i nastroje w czasie

B. MODEL PRZYCZYNOWY SCM / DBN / WORLD MODEL

- Graf przyczynowy G
- Strukturalny model przyczynowy (SCM) $x = f_x(pa_x, e_x)$
- Dynamika świata (DBN / world model) $p(s_{t+1}, z_{t+1} \mid s_t, a_t, c_t)$
- Do usuwania skrótów korelacyjnych i wnioskowania kontrfaktycznego

6. BRAMKA EPISTEMICZNA (BEZPIECZEŃSTWO I POKORA)

czy decyzja jest wystarczająco pewna i bezpieczna?

$$u_t^{epi} \leq \theta_u \quad \wedge \quad \rho_t \leq \theta_r \quad \wedge \quad P(s_t \in \Gamma_\alpha) \geq 1 - \alpha$$

oraz spełnienie ograniczeń bezpieczeństwa i etyki

TAK

NIE

7A. PLANOWANIE (PLANNER)

wybór reakcji w horyzoncie H

$$\pi^* = \arg\max_\pi \mathrm{E}\left[\sum_{k=0}^{H} \gamma^k R(s_{t+k}, a_{t+k})\right]$$

Nagroda: adekwatność emocjonalna + skuteczność + bezpieczeństwo - koszt

8A. DZIAŁANIE (ODPOWIEDŹ SYSTEMU)

dialog / ruch / interwencja / wsparcie

7B. DOPYTANIE / ABSTENCJA / ESKALACJA

gdy niepewność zbyt wysoka lub ryzyko zbyt duże

Dopytanie	Abstencja	Eskalacja
pytanie doprecyzowujące	brak decyzji / wstrzymanie działania	przekazanie do człowieka (operator / opiekun)

nowe dane zamykają pętle

NOTACJE

$x_t = (x_t^a, x_t^v, x_t^l, x_t^s)$ — obserwacja multimodalna

- z_t — reprezentacja z fuzji
- h_t — historia dialogu
- u_t^{usr} — model użytkownika
- c_t — kontekst sytuacyjny
- s_t — stan użytkownika ukryty
- $b_t(s)$ — belief state (rozkład)
- $\Gamma_\alpha(x_t)$ — zbiór predykcji conformal
- u_t^{epi}, u_t^{ale} — niepewności
- ρ_t — ryzyko decyzyjne

PRZYKŁADOWA FUNKCJA NAGRODY R

$$R = w_1 R_{emo} + w_2 R_{task} + w_3 R_{soc} - w_4 R_{risk} - w_5 R_{cost}$$

- R_{emo} adekwatność emocjonalna
- R_{task} skuteczność zadaniowa
- R_{soc} harmonia społeczna
- R_{risk} bezpieczeństwo / ryzyko
- R_{cost} koszt (czas, wysiłek, zasoby)

GWARANCJE I KORZYŚCI

- Gwarancja pokrycia (conformal)
- Redukcja błędów potwierdzenia dzięki korekcie przyczynowej
- Jawna ocena ryzyka i niepewności
- Epistemiczna pokora → dopytanie zamiast pochopnej decyzji
- Decyzje optymalizują adekwatność emocjonalną, nie tylko cel zadaniowy

LEGENDA (PRZEPŁYW)

- Przepływ główny
- Przepływ informacji / kontekstu
- Pętla danych (nowe dane)
- Percepcja i wnioskowanie
- Decyzje i bramka
- Działania
- Pamięć i wiedza

Autorski framework: SAPIENS

Socially-aware, probalistic, intent-aware, epistemically-calibrated, neuro-symbolic system

