

Eryka PROBIERZ

Uczelnia Techniczno-Handlowa im. H. Chodkowskiej w Warszawie

HYBRYDOWE SYSTEMY DECYZYJNE W ROBOTYCE SPOŁECZNEJ: WIELOWARSTWOWA ARBITRACJA AUTONOMII W UKŁADACH WSPÓLDZIELONEJ KONTROLI

Streszczenie. W artykule przedstawiono rozszerzoną koncepcję hybrydowego systemu decyzyjnego dla robota społecznego, ujmowanego jako interakcyjny układ automatyki dyskretnej. Zaproponowany model formalizuje współdzielenie kontroli między człowiekiem a agentem autonomicznym jako problem arbitrażu autonomii realizowanego przez warstwę nadzorczą. W przeciwieństwie do prostych reguł przełączania uwzględniono tu zagregowany poziom ryzyka, wiarygodność percepcji i interpretacji intencji, obciążenie użytkownika, zgodność zachowania robota z normami społecznymi, zaufanie operacyjne oraz możliwość jawnego przejęcia kontroli. Opracowano formalny model stanu, funkcje oceny ryzyka i pewności, funkcje użyteczności dla trybów pracy oraz mechanizm histerezy ograniczający oscylacje między stanami. Architektura została opisana blokowo, a logikę arbitrażu przedstawiono w postaci diagramu decyzyjnego, tabel przejść i referencyjnej implementacji w języku Python. Dodatkowo przeprowadzono syntetyczną walidację scenariuszową obejmującą 60 przebiegów po 240 kroków decyzyjnych. Wyniki wskazują, że model rozszerzony zmniejsza liczbę przełączeń trybów i podnosi konserwatywność decyzji w sytuacjach niepewności, co jest pożądane zwłaszcza w robotach usługowych, edukacyjnych i opiekuńczych. Opracowanie ma charakter koncepcyjno-symulacyjny i stanowi podstawę do dalszej kalibracji na danych z rzeczywistych interakcji człowiek-robot.

HYBRID DECISION-MAKING SYSTEMS IN SOCIAL ROBOTICS: MULTILAYER AUTONOMY ARBITRATION IN SHARED-CONTROL ARCHITECTURES

Summary. The paper develops an extended hybrid decision-making framework for a social robot treated as an interactive discrete automation system. Shared control is formalized as autonomy arbitration performed by a supervisory layer that combines aggregated risk, perception and intention confidence, user workload, social-norm compliance, operational trust, and explicit human override. The study proposes a state model, utility functions for operating modes, transition guards, and a hysteresis mechanism limiting unstable switching. The concept is complemented by a reference Python implementation and a synthetic scenario-based validation. The results indicate that the extended model reduces mode-switching frequency and increases decision conservativeness under uncertainty, which is desirable for social robots operating near humans in assistive, educational, and service settings.

1. Wprowadzenie

Roboty społeczne są obecnie rozwijane jako systemy cyber-fizyczno-społeczne, w których klasyczne sterowanie ruchem jest ściśle związane z percepcją człowieka, interpretacją kontekstu oraz generowaniem zachowań społecznie adekwatnych. Najnowsze przeglądy systematyczne pokazują, że społecznie asystujące roboty są wdrażane w edukacji, opiece, usługach i środowiskach współpracy, lecz wraz ze wzrostem autonomii rosną również wymagania dotyczące bezpieczeństwa, akceptowalności i przewidywalności działania [9]. Z punktu widzenia automatyki nie wystarcza już optymalizacja trajektorii ani klasyczne sterowanie wykonawcze. W pętli sprzężenia zwrotnego pojawia się użytkownik wraz z jego intencjami, ograniczeniami poznawczymi, zaufaniem, preferencjami oraz reakcjami na zachowanie robota. W konsekwencji problem sterowania powinien być formułowany nie tylko w kategoriach realizacji zadania, ale także jako utrzymywanie odpowiedniego poziomu sprawczości człowieka i kontroli nad zakresem autonomii systemu [5], [8]. To przesunięcie akcentu jest szczególnie istotne w robotyce społecznej, gdzie błędna decyzja może mieć konsekwencje nie tylko fizyczne, lecz również relacyjne: naruszenie dystansu interpersonalnego, nieadekwatny moment wejścia w interakcję, zbyt duża inicjatywa robota albo zbyt częste domaganie się potwierdzeń. Dlatego coraz częściej postuluje się wprowadzanie warstw nadzorczych, które arbitrują poziom autonomii, zamiast pozostawiać decyzję wyłącznie agentowi planującemu zachowanie [4], [6]. Motywacją niniejszego opracowania jest potrzeba formalnego opisu warstwy nadzorczej odpowiedzialnej za arbitraż autonomii. W artykule rozwinięto tę koncepcję do postaci modelu o zdefiniowanych zmiennych stanu, funkcjach oceny, warunkach przejść i tabelach parametrów.

2. Problem badawczy i wkład opracowania

Literatura dotycząca współdzielonej autonomii jest rozbudowana, jednak wiele prac koncentruje się na jednym wymiarze problemu: nawigacji, planowaniu zachowania, ocenie zaufania albo wyjaśnialności decyzji. Brakuje natomiast zwartych modeli decyzyjnych, które w jednym formalizmie łączą: ryzyko fizyczne i społeczne, pewność percepcji, obciążenie użytkownika, zgodność z normami społecznymi, poziom zaufania oraz możliwość natychmiastowego przejęcia sterowania przez człowieka [2], [3], [7], [12]. Problem badawczy można więc sformułować następująco: jak zaprojektować dyskretny kontroler nadzorczy dla robota społecznego, który będzie wybierał między trybem ręcznym, współdzielonym, autonomicznym, trybem żądania potwierdzenia i zatrzymaniem bezpieczeństwa, przy jednoczesnym respektowaniu priorytetów bezpieczeństwa i społecznej adekwatności interakcji?

Wkład niniejszego artykułu ma cztery poziomy. Po pierwsze, zaproponowano wektor stanu opisujący sytuację decyzyjną w sposób wystarczająco złożony dla robotyki społecznej. Po drugie, sformalizowano zagregowane miary ryzyka i pewności. Po trzecie, wprowadzono funkcje użyteczności dla poszczególnych trybów pracy oraz mechanizm histerezy, który redukuje zjawisko oscylacji między stanami. Po czwarte, model zweryfikowano w syntetycznej symulacji scenariuszowej oraz uzupełniono tabelami parametrów i przejść, ułatwiającymi późniejszą implementację.

3. Założenia modelowe i formalizacja problemu

3.1. Wektor stanu i zmienne decyzyjne

Niech stan sytuacji decyzyjnej w chwili t będzie opisany wektorem:

$$x(t) = [\rho(t), \kappa(t), W(t), \tau(t), s(t), o(t), z(t)] \quad (1)$$

gdzie: $\rho(t)$ oznacza zagregowany poziom ryzyka, $\kappa(t)$ - łączną pewność percepcji i interpretacji sytuacji, $W(t)$ - obciążenie użytkownika, $\tau(t)$ - estymatę zaufania operacyjnego, $s(t)$ - zgodność zachowania robota z normami społecznymi i lokalnym kontekstem interakcji, $o(t)$ - sygnał jawnego przejścia kontroli przez człowieka, natomiast $z(t)$ - wskaźnik konfliktu między rekomendacją robota a preferencją użytkownika. Przyjęto, że zmienne $\rho(t)$, $\kappa(t)$, $W(t)$, $\tau(t)$ i $s(t)$ są znormalizowane do przedziału $\langle 0,1 \rangle$. Takie ujęcie ułatwia późniejszą agregację sygnałów pochodzących z różnych kanałów, np. z analizy obrazu, rozpoznawania mowy, klasyfikacji intencji, pomiaru opóźnień reakcji operatora czy prostych heurystyk interfejsowych. W praktyce wartości te nie muszą być mierzone bezpośrednio; mogą stanowić wynik filtracji i estymacji na poziomie nadrzędnym [3], [8], [11].

3.2. Agregacja ryzyka i pewności

Zagregowany poziom ryzyka został zapisany jako ważona suma trzech składowych: ryzyka fizycznego, ryzyka społecznego oraz ryzyka kontekstowego.

$$\rho(t) = 0,50 r_{phys}(t) + 0,30 r_{soc}(t) + 0,20 r_{ctx}(t) \quad (2)$$

Składowa r_{phys} odnosi się do potencjalnej kolizji, naruszenia stref bezpieczeństwa i niepożądanych kontaktów fizycznych. Składowa r_{soc} obejmuje możliwość wywołania dyskomfortu, naruszenia norm proksemicznych, nieadekwatnej inicjacji kontaktu lub przerwania aktywności użytkownika. Składowa r_{ctx} reprezentuje ryzyko wynikające z cech sytuacji, np. wieloosobowej interakcji, ograniczonej przestrzeni, obecności osób wrażliwych lub wysokiej niejednoznaczności zadania [4], [6], [7].

Analogicznie zdefiniowano zagregowaną pewność percepcyjno-decyzyjną:

$$\kappa(t) = 0,45 c_{perc}(t) + 0,35 c_{int}(t) + 0,20 c_{ctx}(t) \quad (3)$$

gdzie c_{perc} oznacza wiarygodność percepcji sensorycznej, c_{int} - pewność interpretacji intencji człowieka, a c_{ctx} - spójność bieżącej sytuacji z modelem kontekstowym robota. Takie rozbieżenie jest uzasadnione tym, że błąd decyzyjny nie musi wynikać z jednego źródła; równie problematyczne mogą być: błędna segmentacja sceny, niewłaściwie rozpoznana intencja oraz brak zgodności między obserwowanym zachowaniem a typowym przebiegiem interakcji [1], [9], [14]. Zmienne modelu zostały przedstawione w tabeli 1.

Tabela 1

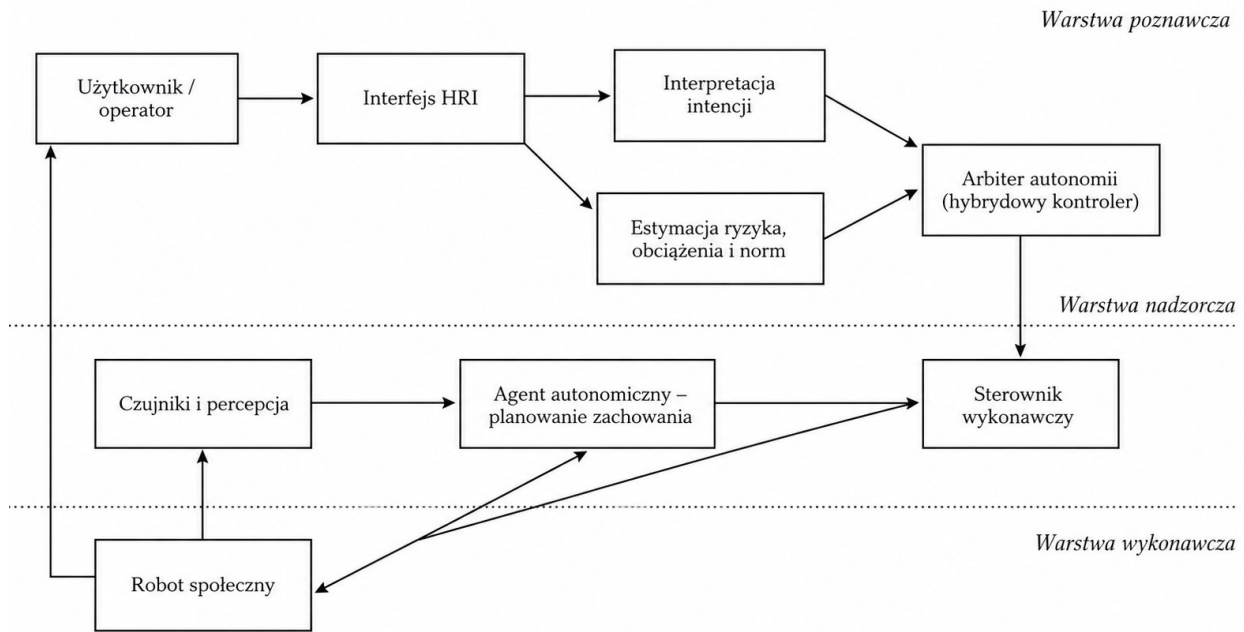
Zmienne modelu i ich interpretacja

Symbol	Znaczenie	Zakres	Źródło estymacji	Rola w arbitrażu
$\rho(t)$	zagregowane ryzyko	0-1	fuzja ryzyka fizycznego, społecznego i kontekstowego	blokada AUTO, aktywacja STOP
$\kappa(t)$	łączna pewność percepcji i interpretacji	0-1	moduł percepcji, klasyfikator intencji, ocena spójności	warunek AUTO i SHARED
$W(t)$	obciążenie użytkownika	0-1	tempo reakcji, złożoność zadania, wskaźniki interfejsowe	uzasadnienie wspomagania autonomicznego
$\tau(t)$	zaufanie operacyjne	0-1	historia interakcji, oceny ex post, zgodność działania z oczekiwaniami	modyfikacja preferencji AUTO/SHARED
$s(t)$	zgodność społeczna	0-1	reguły norm społecznych, proksemika, analiza sytuacji	ograniczenie decyzji przy konflikcie norm
$o(t)$	jawne przejęcie kontroli	{0,1}	interfejs HRI, przycisk awaryjny, komenda głosowa	priorytet trybu MANUAL
$z(t)$	konflikt preferencji	{0,1}	rozbieżność między rekomendacją systemu a deklaracją użytkownika	wymuszenie CONFIRM

4. Architektura hybrydowego systemu decyzyjnego

Architektura ma charakter warstwowy. Warstwa percepcyjno-poznawcza przetwarza dane z czujników i interfejsu HRI, warstwa nadzorcza arbitruje poziom autonomii, a warstwa wykonawcza realizuje zachowanie robota. Takie rozdzielenie pozwala odizolować logikę bezpieczeństwa i społecznej adekwatności od algorytmów niskiego poziomu odpowiedzialnych za napęd, lokomocję lub gestykulację [1], [4], [12]. Kluczową rolę odgrywa arbiter autonomii. Nie zastępuje on planera zachowania, lecz decyduje, czy rekomendacja agenta autonomicznego może być wykonana bezpośrednio, powinna zostać uzgodniona z człowiekiem, czy też musi zostać zablokowana. Z perspektywy projektowej jest to warstwa o charakterze sterowania nadzorczego, nadrzędna wobec wykonawczego sterowania ruchem.

Znaczenie takiej architektury potwierdzają badania nad społecznie świadomą nawigacją oraz normatywnym sterowaniem robotów. Współczesne przeglądy wskazują, że redukcja dyskomfortu człowieka wymaga łączenia bezpieczeństwa fizycznego, bezpieczeństwa postrzeganego, naturalności ruchu i zgodności z normami społecznymi [7]. Z kolei architektury takie jak SONAR pokazują, że świadomość norm społecznych może być implementowana jako jawny moduł sterowania, a nie wyłącznie jako efekt uboczny uczenia modelu [4]. Rysunek 1 przedstawia architekturę systemu.



Rys. 1. Warstwowa architektura hybrydowego systemu współdzielenia kontroli.

Źródło: opracowanie własne.

5. Model arbitrażu autonomii jako automat hybrydowy

Zbiór dopuszczalnych trybów pracy przyjęto w postaci $M = \{\text{MANUAL}, \text{SHARED}, \text{AUTO}, \text{CONFIRM}, \text{STOP}\}$. W odróżnieniu od prostego automatu stanów przejścia nie są definiowane wyłącznie przez pojedyncze progi, lecz przez kombinację warunków bezpieczeństwa i funkcji użyteczności. Rozwiązanie to zachowuje interpretowalność modelu, a jednocześnie pozwala uwzględniać wielokryterialność decyzji. Najwyższy priorytet ma bezpieczeństwo. Jeżeli $\rho(t) \geq \rho_{stop}$, system przechodzi do stanu STOP niezależnie od pozostałych zmiennych. Drugi priorytet ma jawna sprawczość człowieka: jeżeli $o(t) = 1$, uruchamiany jest tryb MANUAL. Dopiero po spełnieniu tych warunków system ocenia niepewność i konflikt społeczny; jeśli $\kappa(t) < \kappa_{min}$ albo $s(t) < s_{min}$ lub $z(t) = 1$, aktywowany jest tryb CONFIRM. W przeciwnym razie rozstrzygnięcie zapada między trybami SHARED i AUTO.

Do rozróżnienia między SHARED i AUTO wprowadzono funkcje użyteczności:

$$J_{AUTO} = 0,30 \kappa + 0,18 W + 0,14 \tau + 0,12 s - 0,38 \rho \quad (4)$$

$$J_{SHARED} = 0,25 \kappa + 0,16 (1 - |W - 0,55|) + 0,18 s + 0,16 \tau - 0,12 \rho \quad (5)$$

Tryb AUTO wybierany jest tylko wtedy, gdy $J_{AUTO} > J_{SHARED} + \Delta$ oraz równocześnie spełnione są dodatkowe warunki progowe: $\kappa(t) \geq \kappa_{auto}$, $\rho(t) < \rho_{auto}$, $s(t) \geq s_{auto}$ i $\tau(t) \geq \tau_{auto}$. Wprowadzenie marginesu Δ oraz minimalnego czasu przebywania w stanie stanowi mechanizm histerezy ograniczający przełączenia wymuszone krótkotrwałymi fluktuacjami sygnałów. Takie sformułowanie modelu jest spójne z wynikami badań wskazujących, że poziom autonomii powinien zależeć nie tylko od skuteczności zadaniowej, ale także od wyjaśnialności, postrzeganej kompetencji i zgodności działania robota z oczekiwaniami człowieka [2], [3], [5]. W tabeli 2 przedstawiono parametry arbitrażu autonomii, natomiast w tabeli 3 logikę przejść między trybami pracy. Rysunek 2 przedstawia diagram decyzyjny arbitra autonomii.

Tabela 2

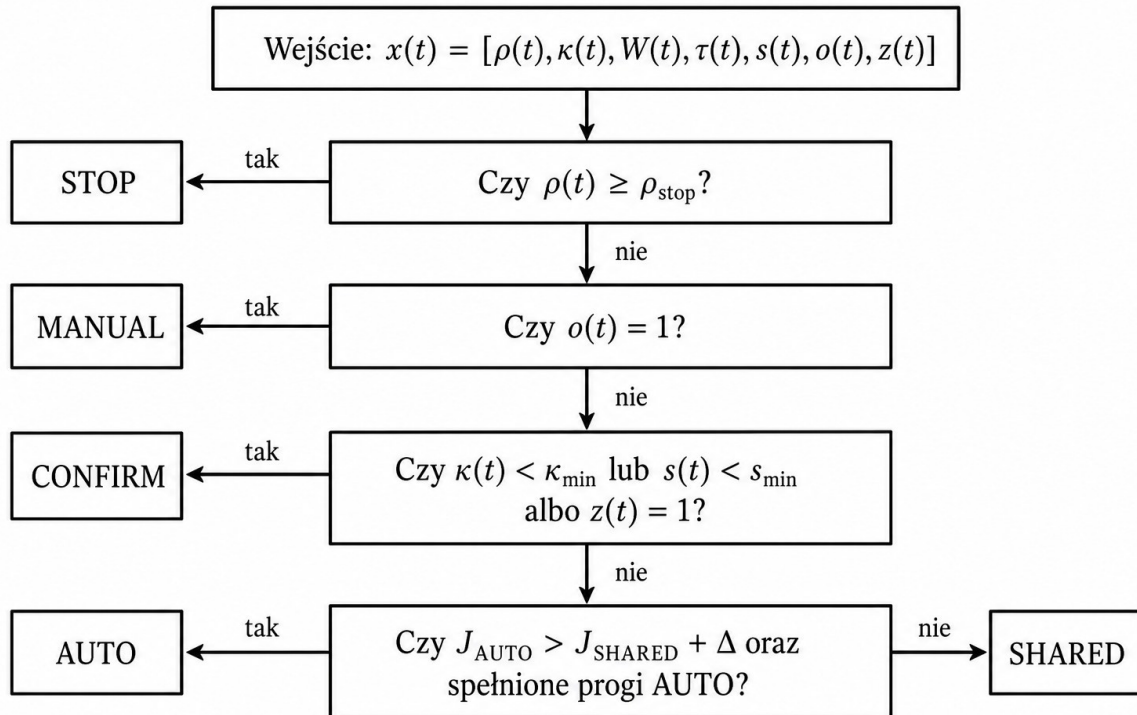
Parametry arbitrażu autonomii zastosowane w modelu referencyjnym

Parametr	Wartość	Znaczenie	Uwagi implementacyjne
ρ_{stop}	0,62	próg zatrzymania bezpieczeństwa	aktywny niezależnie od pozostałych kryteriów
κ_{min}	0,58	minimalna akceptowalna pewność	poniżej progu wymagane CONFIRM
s_{min}	0,60	minimalna zgodność społeczna	zabezpiecza przed nieadekwatnym zachowaniem
ρ_{auto}	0,45	maksymalny poziom ryzyka dla AUTO	AUTO tylko przy ryzyku niskim
κ_{auto}	0,72	wymagana pewność dla AUTO	chroni przed nadmierną autonomią
s_{auto}	0,67	minimalna zgodność społeczna dla AUTO	wymaga wysokiej adekwatności interakcji
τ_{auto}	0,55	minimalne zaufanie operacyjne	wzmacnia rolę historii interakcji
Δ	0,025	margines preferencji J_{AUTO} nad J_{SHARED}	eliminuje wybór przy niemal remisie
d_{min}	2 próbki	minimalny czas pozostania w stanie	redukuje oscylacje przełączeń

Tabela 3

Logika przejść między trybami pracy

Przejście	Warunek strażniczy	Priorytet	Interpretacja
* $\rightarrow$ STOP	$\rho(t) \geq \rho_{stop}$	1	eskalacja bezpieczeństwa; natychmiastowe zatrzymanie
* $\rightarrow$ MANUAL	$o(t) = 1$ oraz $\rho(t) < \rho_{stop}$	2	człowiek jawnie przejmuje kontrolę
SHARED/AUT O $\rightarrow$ CONFIRM	$\kappa(t) < \kappa_{min}$ lub $s(t) < s_{min}$ lub $z(t) = 1$	3	niepewność, konflikt lub niezgodność społeczna
SHARED $\rightarrow$ AUTO	$J_{AUTO} > J_{SHARED} + \Delta$ i spełnione progi AUTO	4	autonomia dopuszczona jako wspomaganie
AUTO $\rightarrow$ SHARED	spadek $\tau(t)$ lub wzrost $\rho(t)$, przy niespełnieniu warunków AUTO	4	powrót do współdzielenia kontroli
CONFIRM $\rightarrow$ SHARED	potwierdzenie użytkownika przy warunkach bezpiecznych	5	deeskalacja po uzgodnieniu działania
MANUAL $\rightarrow$ SHARED	zwolnienie kontroli i brak stanu awaryjnego	5	powrót do trybu bazowego



Rys. 2. Diagram decyzyjny arbitra autonomii. Źródło: opracowanie własne.

6. Walidacja scenariuszowa i wyniki symulacji

Walidację przeprowadzono w dwóch krokach:

1. Najpierw sprawdzono odpowiedź kontrolera na zestaw jawnie zdefiniowanych scenariuszy reprezentujących typowe sytuacje interakcji człowiek-robot.
2. Następnie wykonano syntetyczną symulację czasową obejmującą 60 przebiegów po 240 kroków decyzyjnych, co łącznie dało 14 400 instancji wyboru trybu pracy. Zmiennym wejściowym nadano ciągłość czasową i okazjonalne zaburzenia, aby odzwierciedlić realistyczne fluktuacje percepcji, ryzyka i obciążenia użytkownika. Porównano model bazowy, odpowiadający prostszej logice regułowej, z modelem rozszerzonym opisanym w niniejszym artykule. Analiza nie ma charakteru empirycznego badania z udziałem użytkowników; jej celem jest wewnętrzna walidacja spójności reguł arbitrażu, a nie wykazywanie przewagi jednego rozwiązania w sensie klinicznym lub eksploatacyjnym.

6.1 Procedura generowania scenariuszy syntetycznych

Scenariusze syntetyczne wygenerowano w celu sprawdzenia wewnętrznej spójności logiki arbitrażu autonomii oraz porównania modelu bazowego z modelem rozszerzonym. Tabela 4 przedstawia scenariusze referencyjne wraz z decyzjami modelu rozszerzonego. Dla każdego z 60 przebiegów symulacyjnych wygenerowano sekwencję 240 kroków decyzyjnych. Zmienne ciągłe $\rho(t)$, $\kappa(t)$, $W(t)$, $\tau(t)$ oraz $s(t)$ generowano jako ograniczone do przedziału $\langle 0,1 \rangle$ szeregi czasowe z łagodną zależnością od poprzedniego

kroku, tak aby odzwierciedlić stopniowe zmiany ryzyka, obciążenia, zaufania i jakości percepcji. Dla zmiennych ciągłych przyjęto rozkłady normalne obcięte do przedziału $\langle 0,1 \rangle$. Zagregowane ryzyko $\rho(t)$ generowano wokół wartości średniej 0,30, pewność percepcyjno-decyzyjną $\kappa(t)$ wokół wartości 0,76, obciążenie użytkownika $W(t)$ wokół wartości 0,55, zaufanie operacyjne $\tau(t)$ wokół wartości 0,68, natomiast zgodność społeczną $s(t)$ wokół wartości 0,72. W wybranych krokach wprowadzano krótkotrwałe zaburzenia, obejmujące wzrost ryzyka, spadek pewności percepcji, spadek zgodności społecznej lub wzrost obciążenia użytkownika. Zaburzenia te miały charakter epizodyczny i trwały od kilku do kilkunastu kroków decyzyjnych.

Sygnał jawnego przejścia kontroli przez człowieka $o(t)$ generowano jako zmienną binarną. Wartość $o(t)=1$ oznaczała aktywację trybu ręcznego przez użytkownika, np. za pomocą interfejsu HRI, komendy głosowej lub przycisku awaryjnego. Konflikt preferencji $z(t)$ również modelowano jako zmienną binarną. Wartość $z(t)=1$ oznaczała rozbieżność między rekomendacją robota a preferencją użytkownika lub wystąpienie sytuacji wymagającej dodatkowego potwierdzenia działania.

Model bazowy zdefiniowano jako uproszczony arbiter regułowy, który wykorzystywał priorytet zatrzymania bezpieczeństwa, jawne przejście kontroli oraz prosty wybór między trybami AUTO i SHARED na podstawie porównania funkcji użyteczności. W modelu bazowym nie stosowano marginesu Δ , minimalnego czasu pozostawiania w stanie ani dodatkowych progów dopuszczalności trybu AUTO, takich jak κ_{AUTO} , ρ_{AUTO} , s_{AUTO} i τ_{AUTO} . Model rozszerzony odpowiadał pełnej logice arbitrażu opisanego w rozdziale 5.

Dla zachowania porównywalności oba modele testowano na tym samym zestawie wygenerowanych danych wejściowych. Oznacza to, że dla modelu bazowego i rozszerzonego zastosowano identyczne przebiegi zmiennych $\rho(t)$, $\kappa(t)$, $W(t)$, $\tau(t)$, $s(t)$, $o(t)$ oraz $z(t)$. W celu zapewnienia powtarzalności eksperymentu zastosowano stałe ziarno generatora liczb pseudolosowych.

Wyniki przedstawione w tabeli 5 są średnimi z 60 przebiegów symulacyjnych.

Tabela 4

Scenariusze referencyjne i decyzje modelu rozszerzonego

Scenariusz	$\rho(t)$	$\kappa(t)$	$W(t)$	$\tau(t)$	$s(t)$	$o(t)$	$z(t)$	Tryb wynikowy
Rutynowa asysta informacyjna	0,20	0,86	0,42	0,76	0,85	0	0	SHARED
Silne obciążenie użytkownika przy bardzo korzystnych warunkach	0,08	0,97	1,00	0,99	0,99	0	0	AUTO
Niska pewność interpretacji intencji	0,20	0,51	0,50	0,66	0,78	0	0	CONFIRM
Konflikt norm społecznych w ciasnej przestrzeni	0,33	0,80	0,55	0,62	0,46	0	1	CONFIRM

Jawne przejęcie kontroli przez operatora	0,20	0,81	0,44	0,74	0,80	1	0	MANUAL
Nagły wzrost ryzyka kolizyjnego	0,64	0,84	0,58	0,70	0,76	0	0	STOP
Spadek zaufania po błędnej rekomendacji	0,24	0,76	0,60	0,32	0,74	0	0	SHARED
Wielostanowiskowa interakcja grupowa	0,30	0,72	0,68	0,60	0,66	0	0	SHARED

Objaśnienie: $o(t)=1$ oznacza jawne przejęcie kontroli przez użytkownika, natomiast $z(t)=1$ oznacza konflikt preferencji lub konieczność dodatkowego uzgodnienia działania.

Tabela 5

Porównanie modelu bazowego i rozszerzonego w symulacji syntetycznej

Metryka	Model bazowy	Model rozszerzony
Liczba przełączeń na 100 kroków	18,4	15,0
Udział trybu AUTO [%]	53,5	12,8
Udział trybu SHARED [%]	38,0	71,8
Udział trybu CONFIRM [%]	3,1	9,9
Udział trybu MANUAL [%]	5,3	5,3
Udział trybu STOP [%]	0,2	0,2
Udział niebezpiecznych decyzji AUTO [%]	0,0	0,0
Udział potwierdzeń nieuzasadnionych [%]	0,0	4,4

Symulacja potwierdziła, że dodanie histerezy zmniejsza częstotliwość przełączeń, a rozszerzenie modelu o zgodność społeczną i zaufanie prowadzi do bardziej zachowawczego arbitrażu. Udział trybu AUTO spadł z 53,5% do 12,8%, natomiast udział trybu SHARED wzrósł do 71,8%. Oznacza to, że model rozszerzony nie traktuje autonomii jako ustawienia domyślnego, lecz jako stan warunkowo dopuszczalny tylko przy bardzo korzystnej konfiguracji zmiennych. Jednocześnie wzrost udziału trybu CONFIRM do 9,9% wskazuje na koszt tej konserwatywności. W środowiskach opiekuńczych lub edukacyjnych taki kompromis może być pożądanym, ponieważ zmniejsza ryzyko nieadekwatnej inicjatywy robota. W zastosowaniach o większej presji wydajnościowej progi arbitrażu należałoby jednak kalibrować adaptacyjnie, zależnie od profilu zadania i doświadczenia użytkownika [8], [11], [13].

7. Dyskusja

Zaproponowany model łączy zalety systemów regułowych i modeli użyteczności. Z jednej strony zachowuje przejrzystość i audytowalność – można wskazać, dlaczego system zablokował autonomię lub zażądał potwierdzenia. Z drugiej strony nie

sprowadza decyzji do pojedynczego progu, lecz uwzględnia kompromis między bezpieczeństwem, potrzebą wsparcia użytkownika i społeczną adekwatnością działania. Takie podejście pozostaje zgodne z kierunkiem badań nad wyjaśnialnymi i adaptacyjnymi architekturami HRC. W ostatnich latach podkreśla się, że zaufanie nie jest pochodną samej skuteczności zadaniowej, lecz wynika z połączenia kompetencji, przewidywalności, komunikowalności intencji oraz zgodności zachowania z oczekiwaniami człowieka [2], [3], [5]. W robotyce społecznej ta zależność jest jeszcze silniejsza, ponieważ nawet technicznie poprawna akcja może zostać oceniona negatywnie, jeśli zostanie wykonana w niewłaściwym momencie lub bez adekwatnego sygnału uprzedzającego.

Ograniczeniem modelu jest sposób estymacji zmiennych ukrytych, zwłaszcza obciążenia użytkownika i zaufania operacyjnego. W praktyce ich pomiar wymagałby łączenia wskaźników behawioralnych, historii interakcji oraz danych z interfejsu. W wielu zastosowaniach niepożądane byłoby jednak nadmiernie inwazyjne monitorowanie człowieka. Dlatego dalsze badania powinny zmierzać ku lekkim, nieinwazyjnym estymatorom, wykorzystującym czas odpowiedzi, liczbę korekt poleceń, częstotliwość odwołań i jawne oceny użytkownika. Drugim ograniczeniem jest brak walidacji z udziałem uczestników. W syntetycznej symulacji można zweryfikować wewnętrzną logikę sterownika, ale nie da się na jej podstawie przesądzić, czy użytkownicy rzeczywiście odczytują częstsze przejścia do trybu CONFIRM jako przejaw bezpieczeństwa, czy raczej jako uciążliwość. Wyniki badań nad grupową HRI i generowaniem zachowania społecznego sugerują, że percepcja naturalności i adekwatności może różnić się zależnie od liczby uczestników, dynamiki zadania i lokalnych norm interakcji [10], [14].

Zaletą modelu jest natomiast jego kompatybilność z uczeniem maszynowym. Agent autonomiczny może wykorzystywać modele głębokie do percepcji, rozpoznawania intencji lub planowania zachowania, podczas gdy warstwa nadzorcza zachowuje odpowiedzialność za dopuszczenie lub zablokowanie autonomii. Taki podział jest szczególnie użyteczny w systemach wysokiego ryzyka, gdzie pełna autonomia bez warstwy ochronnej mogłaby prowadzić do decyzji trudnych do wyjaśnienia i audytu.

8. Wnioski

Hybrydowe systemy decyzyjne w robotyce społecznej można efektywnie opisywać jako dyskretnie układy nadzorcze, które arbitrują poziom autonomii między człowiekiem a agentem autonomicznym. Zaproponowany model rozszerza klasyczne podejście regułowe o formalizację stanu, funkcje oceny, tabele parametrów oraz mechanizm wyboru trybu pracy. Najważniejszym wnioskiem jest to, że autonomia w robotyce społecznej nie powinna być traktowana jako wartość domyślna. Jej uruchomienie musi być warunkowane niskim ryzykiem, wysoką pewnością percepcji, odpowiednią zgodnością społeczną oraz wystarczającym poziomem zaufania operacyjnego. W przeciwnym razie system powinien preferować kontrolę współdzieloną, żądanie potwierdzenia albo zatrzymanie bezpieczeństwa. Przyszłe badania powinny objąć eksperymenty z udziałem użytkowników, kalibrację wag i progów arbitrażu, włączenie kanału wyjaśnień generowanych w czasie rzeczywistym

oraz ocenę wpływu modelu na płynność interakcji, poczucie sprawczości i akceptację robota. Warto także porównać przedstawiony automat hybrydowy z podejściami probabilistycznymi i uczącymi się, zachowując jednak nadrzędne wymagania bezpieczeństwa i audytowalności decyzji.

FINANSOWANIE I DANE. Praca została zrealizowana przy wsparciu finansowym Uczelni Techniczno-Handlowej im. Heleny Chodkowskiej w ramach dotacji na prace badawcze w roku akademickim 2025/2026. Repozytorium zawierające kod źródłowy modelu, skrypty symulacyjne oraz dokumentację jest dostępne pod adresem: <https://github.com/Rochena/HAA---Hybrid-Autonomy-Arbitration> [dostęp: 30.06.2026r.]

LITERATURA

1. Asprino L., Ciancarini P., Nuzzolese A.G., Presutti V., Russo A.: A reference architecture for social robots. *Journal of Web Semantics*, 2022, Vol. 72, art. 100683.
2. Cantucci F., Falcone R., Marini M.: Redefining User Expectations: The Impact of Adjustable Social Autonomy in Human-Robot Interaction. *Electronics*, 2024, Vol. 13, No. 1, art. 127.
3. Cantucci F., Marini M., Falcone R.: Trustworthiness Assessment of an Adaptive and Explainable Robot in a Real Environment. *International Journal of Social Robotics*, 2025, Vol. 17, p. 2199-2210.
4. Dell'Anna D., Jamshidnejad A.: SONAR: An Adaptive Control Architecture for Social Norm Aware Robots. *International Journal of Social Robotics*, 2024, Vol. 16, p. 1969-2000.
5. Elgohr A.T., Rashad M., El-Gendy E.M., Shaaban W., Saafan M.M.: Trust as a design principle in human-robot collaboration: a review of explainable and adaptive control. *Artificial Intelligence Review*, 2026, Vol. 59, art. 134.
6. International Organization for Standardization: ISO 13482:2014. Robots and robotic devices — Safety requirements for personal care robots. International Organization for Standardization, Geneva 2014.
7. Karwowski J., Szynekiewicz W., Niewiadomska-Szynekiewicz E.: Bridging Requirements, Planning, and Evaluation: A Review of Social Robot Navigation. *Sensors*, 2024, Vol. 24, No. 9, art. 2794.
8. Liu Y., Caldwell G., Rittenbruch M., Teixeira M.B.F., Burden A., Guertler M.: What Affects Human Decision Making in Human-Robot Collaboration?: A Scoping Review. *Robotics*, 2024, Vol. 13, No. 2, art. 30.
9. Maure R., Bruno B.: Autonomy in socially assistive robotics: a systematic review. *Frontiers in Robotics and AI*, 2025, Vol. 12, art. 1586473.
10. Nigro M., Akinrintoyo E., Salomons N., Spitale M.: Social Group Human-Robot Interaction: A Scoping Review of Computational Challenges. *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2025, p. 468-478.

11. Sobhani M., Giuliani M., Smith J., Pipe A., Peer A.: Evaluating a Decision-Making Architecture in Human-Robot Collaboration Experiments. *International Journal of Social Robotics*, 2025, Vol. 17, p. 2579-2600.
12. Tozadore D.C., Romero R.A.F.: Multiuser design of an architecture for social robots in education: teachers, students, and researchers perspectives. *Frontiers in Robotics and AI*, 2024, Vol. 11, art. 1409671.
13. Udupa S., Kamat V.R., Menassa C.C.: Shared autonomy in assistive mobile robots: a review. *Disability and Rehabilitation: Assistive Technology*, 2023, Vol. 18, No. 6, p. 827-848.
14. Valls Mascaro E., Lee D.: Robot Behavior Generation for Social Human-Robot Interaction. *International Journal of Social Robotics*, 2025, Vol. 17, p. 3211-3230.