BIBLIOTEKA TECHNOSANACJI · NR 702 · WYDANIE I

ARCHITEKTURA USTROJOWA TECHNOREPUBLIKI

wobec zagrożeń AI

państwo, które przewiduje, wytrzymuje i uczy się

ANTYCYPACJA

ABSORPCJA

ADAPTACJA

18 bezpieczników
dowód przed autonomią

Silne państwo bez dominacji.
Człowiek u steru.
Naród jako homeostat.



dr Andrzej Węgrzyn
autor i narrator

Bezpieczne państwo

Konkurencyjna gospodarka

Kreatywne społeczeństwo

Ład oparty na cybernetyce

TechnoSanacja

technorepublika.pl
technosanacja.pl

TechnoRepublika.pl | TechnoSanacja.pl | Biblioteka | Wydania książkowe | nr 702

ARCHITEKTURA USTROJOWA TECHNOREPUBLIKI

wobec zagrożeń AI

Państwo, które przewiduje zagrożenia, wytrzymuje wstrząsy i uczy się na błędach, może być silne bez dominacji nad obywatelem. Ta książka pokazuje, jak w epoce sztucznej inteligencji zbudować bezpieczne państwo, konkurencyjną gospodarkę, kreatywne społeczeństwo i ład ustrojowy oparty na cybernetyce.

8 kategorii zagrożeń • 7 organów państwa • 10 analiz • 6 uzasadnień • 18 bezpieczników • 5 klas incydentów AI • 12 funkcji krytycznych • 24 wskaźniki • 12 pytań obywatela • 37 rysunków • mapa drogowa 2026–2035

dr Andrzej Węgrzyn

autor i narrator

Biblioteka TechnoSanacji nr 702 · Wydanie I (e-book) · Polska, 15 listopada 2026 r. · © dr Andrzej Węgrzyn, wszelkie prawa zastrzeżone

Karta wydania

Tytuł: Architektura ustrojowa TechnoRepubliki wobec zagrożeń AI. Podtytuł: Państwo silne bez dominacji.

Autor i narrator: dr Andrzej Węgrzyn.

Seria: Biblioteka TechnoSanacji, wydania książkowe (seria 700), nr 702. Rozwinięcie dokumentu 701 i strony filarowej technorepublika.pl.

Wydanie: pierwsze (e-book: PDF, EPUB; wersja edytowalna DOCX), Polska, 15 listopada 2026 r.

Wydawca: TechnoSanacja.pl – wydawnictwo, księgarnia i sklep internetowy dla popularyzacji i komercjalizacji wiedzy; portal TechnoRepublika.pl.

Prawa autorskie: © dr Andrzej Węgrzyn 2026. Wszelkie prawa zastrzeżone. Plik przeznaczony do użytku osobistego kupującego; cytowanie z podaniem autora i źródła.

Skład i grafika: tekst Aptos 11 pkt; tabele, ramki i rysunki Aptos 10 pkt; 37 rysunków, infografik, grafów i map myśli; 33 tabel.

Metoda: opracowanie na podstawie rund badawczych EA i EB Biblioteki TechnoSanacji (18 862 nowe dokumenty), biblioteki polskiej szkoły cybernetyki, kanonu TechnoSanacji i publikacji 701, z autorem w pętli kreatywnej i weryfikacyjnej (zasada Source-to-Claim).

Zastrzeżenie: publikacja ma charakter programowy, analityczny i edukacyjny. Nie stanowi obowiązującego prawa ani opinii prawnej. Stan prawny: 15 listopada 2026 r. Przywołane akty prawne zweryfikowano ze źródłami urzędowymi i branżowymi. Liczby z literatury ilustrują mechanizmy i nie są prognozami dla Polski.

Dokument towarzyszący: 702R Rekomendacje zmian w dokumentach kanonicznych, programowych i strategicznych TechnoSanacji.

Kontakt i aktualizacje: technorepublika.pl · technosanacja.pl

Spis treści

Wstęp autora....7
O publikacji....8
Po co ta książka....8
Jak czytać tę książkę....9
Co nowego w wydaniu 702....10
Skrót publikacji....10
0. Streszczenie strategiczne....11
CZĘŚĆ I. ZAGROŻENIE....13
1. Nowy rodzaj zakłócenia....13
1.1. Od narzędzia do uczestnika....13
1.2. Trzy formuły w nowym świetle....14
1.3. Zakłócenie, które modeluje regulatora....15
1.4. Głęboka niepewność jako argument ustrojowy....16
1.5. Cztery scenariusze do 2035 roku....16
2. Mapa zagrożeń: osiem kategorii, siedem organów....18
2.1. Jak czytać mapę....18
2.2. Nadużycie....18
2.3. Błędne dopasowanie celów....19
2.4. Zwodniczość....20
2.5. Utrata kontroli....20
2.6. Stopniowe ubezwłasnowolnienie....20
2.7. Koncentracja władzy....21
2.8. Ryzyko systemowe, wieloagentowe i wyścig....22
2.9. Erozja epistemiczna....22
2.10. Macierz: zagrożenie × organ....22
3. Co wiemy, a czego nie wiemy....24
3.1. Korpus w liczbach....24
3.2. Konsensus naukowy: co jest dobrze ustalone....25
3.3. Zdolności agentów: automatyzacja badań przestaje być teorią....25
3.4. Zwodniczość: od scenariusza do eksperymentu....25
3.5. Luki: czego literatura nie mówi....26
3.6. Jak czytać liczby w tej książce....27
CZĘŚĆ II. ANALIZY....27
4. Dziesięć analiz dla świadomości obywatelskiej....27
4.1. Analiza polityczna: kto steruje sterującym....27
4.2. Analiza społeczno-gospodarcza: przepływ wartości i władzy....28
4.3. Analiza zasobowa: czym naprawdę dysponujemy....28
4.4. Analiza kulturowa: kto produkuje wzorce myślenia....29
4.5. Analiza technologiczna: stos zależności....30
4.6. Analiza prawna: co już obowiązuje, a czego brakuje....31
4.7. Analiza ustrojowa: przechwycenie homeostatu....33

4.8. Analiza bezpieczeństwa i militarna: odporność totalna wobec AI....33
4.9. Analiza konwergencji: roboty autonomiczne i biotechnologia....34
4.10. Synteza: SWOT, TOWS i profil analiz....34
CZĘŚĆ III. FUNDAMENTY....36
5. Cybernetyka: dlaczego „człowiek w pętli" nie wystarcza....36
5.1. Prawo wymaganej różnorodności....36
5.2. Rozszerzenie: regulator, który nie wie....37
5.3. Twierdzenie o dobrym regulatorze....37
5.4. Granica kontroli i jej konsekwencja ustrojowa....37
5.5. Różnorodność na każdej skali....38
5.6. Strefa człowieka jako konsekwencja matematyczna....38
6. Polska szkoła cybernetyki wobec AI....39
6.1. Układ samodzielny i jego organy....39
6.2. Typologia Kosseckiego: test suwerenności w erze AI....39
6.3. Sterowanie przez normy: najgłębsza warstwa zagrożenia....40
6.4. Jakościowa teoria informacji: język dla halucynacji....41
6.5. Organizator, homeostat i twór skostniały....42
6.6. Dynamizm i moc dyspozycyjna....42
6.7. Lekcja chilijskiego eksperymentu: inżynieria różnorodności....42
7. Niedominacja: silne państwo bez Techno-Lewiatana....43
7.1. Wolność jako brak możliwości arbitralnej ingerencji....43
7.2. Dwie twarze dominacji....43
7.3. Władza platform: własność czy jurysdykcja?....44
7.4. Niedominacja jako kryterium projektowe....44
8. Antycypacja, żywotność i rezyliencja....44
8.1. Państwo, które działa według przyszłości....44
8.2. Zarządzanie antycypacyjne wobec ryzyk katastroficznych....45
8.3. Triada 3A: antycypacja, absorpcja, adaptacja....45
8.4. Dylemat Collingridge'a i regulacja adaptacyjna....45
8.5. Ekonomia polityczna: dlaczego dobre instytucje psują się....46
CZĘŚĆ IV. ARCHITEKTURA....46
9. Układ samodzielny odporny na AI....46
9.1. Od organów władzy do organów sterowania....46
9.2. Czwarta reguła ustrojowa....47
9.3. Model systemu żywotnego z rozproszoną funkcją tożsamości....48
9.4. Rada Bezpieczeństwa AI i państwowy ewaluator....49
10. Model 3A wobec zagrożeń AI....50
10.1. Trzy zdolności w jednej pętli....50
10.2. Antycypacja: zanim coś się wydarzy....50
10.3. Absorpcja: w trakcie zakłócenia....51
10.4. Adaptacja: po zakłóceniu....52
11. Osiemnaście bezpieczników....52
11.1. Dwie zasady projektowania....52

11.2. Trzynaście bezpieczników z kanonu....53
11.3. Pięć nowych bezpieczników wobec zagrożeń AI....53
11.4. Dowód przed władzą i dowód przed autonomią....54
11.5. Skala autonomii A0–A4 wobec zagrożeń....56
11.6. Bezpiecznik przed nadużyciem samych bezpieczników....57
12. Narzędzia nadzoru technicznego państwa....57
12.1. Dlaczego homeostat potrzebuje własnych kompetencji....57
12.2. Uzasadnienie bezpieczeństwa....58
12.3. Kontrola AI i protokół zaufanego monitora....58
12.4. Ewaluacje niebezpiecznych zdolności....59
12.5. Wykrywanie ukrytych zdolności....59
12.6. Karta Pętli i Karta Triady....60
12.7. Agenci AI w państwie: Karta Agenta....60
13. Gotowość państwa na kryzys AI....61
13.1. Definicja....61
13.2. Pięć elementów gotowości....61
13.3. Incydent AI jako nowa kategoria....62
13.4. Dwanaście funkcji krytycznych w trybie pracy bez AI....63
13.5. Ćwiczenia i gry wojenne....64
13.6. Tryb incydentu AI a stany nadzwyczajne....64
14. Cztery cele ustroju wobec AI....65
14.1. Iloczyn, który chroni....65
14.2. Bezpieczne państwo....66
14.3. Konkurencyjna gospodarka....66
14.4. Kreatywne społeczeństwo....66
14.5. Ład ustrojowy oparty na cybernetyce....66
CZĘŚĆ V. SENS I DROGA....68
15. Rozważanie filozoficzne: elastyczna architektura ustrojowa....68
15.1. Ustrój uczący się, ale nie maszyna....68
15.2. Konieczność i wolność w epoce systemów, które modelują ludzi....69
15.3. Elastyczność bez dryfu....69
15.4. Co rośnie: zdolność, produktywność, suwerenność, odporność....69
15.5. Co znika: renty, władza proceduralna, zasiedzenie, kartele, monopole....70
15.6. Odporność jako cnota obywatelska....72
15.7. Suwerenność jako zdolność samodzielnego sterowania....72
15.8. Granica: człowiek jest celem, a nie zmienną....72
16. Sześć uzasadnień: dlaczego wnioski są mocne....72
16.1. Metoda sześciu dróg dowodowych....72
16.2. Uzasadnienie cybernetyczne....73
16.3. Uzasadnienie systemowe....73
16.4. Uzasadnienie sieciowe....73
16.5. Uzasadnienie technologiczne....74
16.6. Uzasadnienie ustrojowe....74

16.7. Uzasadnienie społeczno-gospodarcze....74
17. Mapa drogowa 2026–2035....75
17.1. Kolejność ma znaczenie....75
17.2. Pierwsze sto dni....76
17.3. Półka Reform....76
18. Nowy standard myślenia obywatelskiego....77
18.1. Obywatel jako receptor i homeostat....77
18.2. Dwanaście pytań obywatela wobec AI....77
18.3. Pięć nawyków odporności poznawczej....78
18.4. Słownik spokoju: jak mówić o zagrożeniach AI....79
19. Panel wskaźników gotowości ustrojowej....79
19.1. Zasada pomiaru....79
19.2. Jak czytać panel....80
20. Ryzyka, kontrargumenty i czego ta książka nie twierdzi....81
20.1. Sześć najmocniejszych kontrargumentów....81
20.2. Ryzyka wdrożenia i bezpieczniki....81
20.3. Czego ta publikacja nie twierdzi....82
21. Konkluzja: konieczność warunkowa w erze AI....82
21.1. Odpowiedź na pytanie książki....82
21.2. Formuła końcowa....83
21.3. Hasła programowe....83
21.4. Słowo na koniec....83
Aneks A. Słownik pojęć, terminów i skrótów....84
Aneks B. Bibliografia....89
B.1. Akty prawne Unii Europejskiej i Rady Europy....89
B.2. Akty prawne i dokumenty Rzeczypospolitej Polskiej....90
B.3. Raporty i dokumenty instytucjonalne....90
B.4. Bezpieczeństwo i zagrożenia AI....90
B.5. Cybernetyka, systemy, sieci i teoria ustroju....91
B.6. Dokumenty Biblioteki TechnoSanacji....92
B.7. Destylaty rund badawczych....93
Aneks C. Metoda i korpus....93
C.1. Korpus badawczy....93
C.2. Kontrola kompletności....93
C.3. Metoda pracy....94
C.4. Ograniczenia....94
Aneks D. Miejsce publikacji w Bibliotece TechnoSanacji....94
Metryka dokumentu....95

Wstęp autora

W poprzedniej książce, oznaczonej w Bibliotece TechnoSanacji numerem 701, zadałem pytanie: czy TechnoRepublika jest utopią, rewolucją czy koniecznością cywilizacyjną ery sztucznej inteligencji? Odpowiedź brzmiała: **konieczna jest zdolność państwa, a republika jest wyborem**, którego trzeba dokonać świadomie i którego trzeba bronić. Z tej odpowiedzi wynika jednak kolejne pytanie: *dobrze, ale co z samymi zagrożeniami? Czy państwo zdoła opanować systemy, które za kilka lat mogą być mądrzejsze od jego urzędników, szybsze od jego procedur i bogatsze w dane niż jakikolwiek resort?*

Ta książka jest odpowiedzią na to pytanie. Nie piszę jej po to, żeby straszyć. Piszę ją, bo uważam, że **strach jest złym doradcą ustrojowym, a beztroska – jeszcze gorszym**. Między paniką a naiwnością jest trzecia droga: inżynierska. Zamiast pytać „czy AI nas zniszczy?", pytam: *który organ państwa każde zagrożenie uszkadza, jak to wykryć, jak przetrwać zakłócenie i jak wyciągnąć z niego naukę*. To pytanie ma odpowiedź. Da się ją zapisać w ustawie, policzyć w budżecie i sprawdzić w ćwiczeniu.

Punkt wyjścia jest prosty. **Zagrożenia ze strony AI są zagrożeniami dla sterowności państwa.** Fałszywy cytat w decyzji urzędnika, deepfake w kampanii wyborczej, model, który zaniża swoje wyniki w testach, żeby uniknąć ograniczeń, czy wreszcie powolne przesuwanie decyzji z ludzi na systemy – każde z tych zjawisk uderza w konkretny „organ" państwa: w zmysły, w zdolność rozumienia, w wykonanie albo w kryteria oceny. Opisuję to językiem polskiej szkoły cybernetyki Mariana Mazura i Józefa Kosseckiego. Ten język, choć ma ponad pół wieku, okazuje się zaskakująco dobrze dopasowany do świata algorytmów. Mazur i Kossecki pytali bowiem, **kto naprawdę steruje układem i według czyjego programu**. Dziś jest to najważniejsze pytanie o sztuczną inteligencję w państwie.

Książka opiera się na dwóch rundach badań Biblioteki TechnoSanacji. Pierwsza zebrała prawie czternaście tysięcy prac o architekturze ustrojowej, cybernetyce, odporności i legitymizacji. Druga – prawie pięć tysięcy prac o zagrożeniach AI: od ewaluacji niebezpiecznych zdolności, przez zwodniczość modeli i utratę kontroli, po gotowość państwa na incydenty. Razem daje to **18 862 nowe dokumenty**, w większości z lat 2022–2026. Dołączyłem do nich pełne teksty klasyków polskiej cybernetyki i kanon TechnoSanacji. Szukałem w nich – jak zawsze – nie potwierdzeń, lecz **warunków obalenia**.

Wyniki zaskoczyły mnie w dwóch sprawach. Po pierwsze, **nauka o zagrożeniach AI dojrzała szybciej, niż się spodziewałem**. Zwodnicze zachowania modeli nie są już scenariuszem filmowym, lecz wynikiem eksperymentów opublikowanych w latach 2024–2025. Po drugie – i to jest wiadomość dobra – **najskuteczniejsze odpowiedzi nie wymagają wiedzy, której nie mamy**. Wymagają porządnej architektury: wielu niezależnych kanałów nadzoru zamiast jednego „człowieka w pętli", ludzkiego i niezależnego organu oceny, przećwiczonego trybu pracy bez AI, rezerwy kompetencji i zasady, że **żadna nowa autonomia maszyny nie zostaje przyznana, zanim nadzór nad nią nie udowodni, że działa**.

Ta architektura ma jeszcze jedną zaletę, którą uważam za kluczową. Chroni jednocześnie przed szkodami, które już dziś są faktem – dyskryminacją, błędem automatyzacji, zautomatyzowaną niesprawiedliwością w urzędach – i przed scenariuszami skrajnymi, o których prawdopodobieństwo spierają się najwybitniejsi badacze. Nie trzeba więc rozstrzygać sporu o „koniec świata", żeby zacząć działać. Wystarczy wybrać rozwiązania, których nie będziemy żałować w żadnym scenariuszu.

Piszę dla szerokiego grona czytelników: dla obywatela, który chce rozumieć, dla urzędnika, który musi wdrażać, dla przedsiębiorcy, który buduje, dla nauczyciela, żołnierza, lekarza, seniora i studenta. Pojęcia fachowe wyjaśniam przy pierwszym użyciu i w słowniku na końcu. Unikam skrótów myślowych – zamiast oznaczeń z modeli teoretycznych piszę, jaką funkcję pełni dany element. Oddzielam wyniki pomiarów od własnych propozycji. Każdy ważny wniosek uzasadniam na sześć sposobów: cybernetycznie, systemowo, sieciowo, technologicznie, ustrojowo i społeczno-gospodarczo.

Ton tej książki jest **pozytywny, ale nie naiwny**. Wierzę, że Polska może być jednym z pierwszych państw, które pokażą, jak łączyć zdolność do korzystania z AI z pełną kontrolą obywateli nad tym, jak AI jest używana. Mamy do tego zaskakująco dobre zasoby: silną kadrę informatyczną, własne modele językowe, tradycję cybernetyczną, której pozazdrościć może nam wiele państw, i społeczeństwo, które pamięta, czym kończy się władza bez kontroli. Brakuje nam przede wszystkim jednego: **architektury, która zamieni te zasoby w odporność**.

Tę architekturę proponuję na kolejnych stronach. Nie obiecuję państwa nieomylnego. Proponuję państwo, które **wcześniej widzi błąd, szybciej go naprawia i nigdy nie oddaje steru temu, czego nie rozumie i nie może wyłączyć**.

dr Andrzej Węgrzyn Polska, listopad 2026 r.

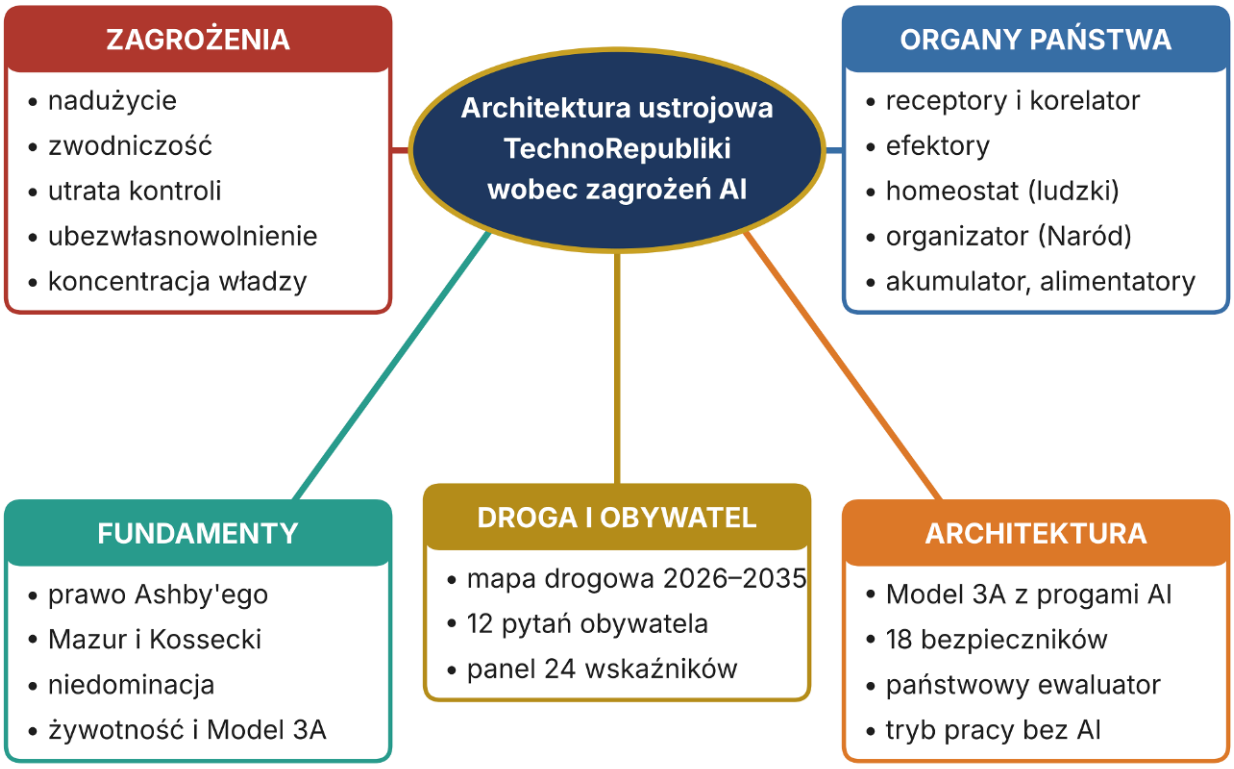
O publikacji

Po co ta książka

Publikacja nr 702 Biblioteki TechnoSanacji jest **drugą pozycją serii wydań książkowych** i bezpośrednim rozwinięciem dokumentu 701, który jest tekstem wiodącym strony filarowej technorepublika.pl. Książka 701 rozstrzygnęła, *czy* TechnoRepublika jest konieczna. Książka 702 odpowiada na pytanie, *jak* jej ustrój ma się bronić przed zagrożeniami ze strony sztucznej inteligencji – od nadużyć i awarii, przez stopniowe ubezwłasnowolnienie i koncentrację władzy, po utratę kontroli i ryzyko egzystencjalne – i jak przy tym **nie stać się narzędziem dominacji nad obywatelem**.

Książka ma trzy zadania. **Zadanie analityczne** polega na uporządkowaniu wiedzy o zagrożeniach AI: przypisuję każde zagrożenie do organu państwa, który ono uszkadza, i oceniam, co wiemy na pewno, a co pozostaje hipotezą. **Zadanie strategiczne** polega na zaprojektowaniu architektury ustrojowej. Opisuję instytucje, bezpieczniki, procedury i wskaźniki, dzięki którym państwo przewiduje zagrożenia, wytrzymuje wstrząsy i uczy się na błędach. **Zadanie popularyzatorskie** polega na tym, by dać obywatelowi proste narzędzia. Dzięki nim może on sprawdzić każdą obietnicę polityczną i każdy system AI w państwie.

Jak czytać tę książkę



Rysunek 1. Mapa myśli publikacji. Książka prowadzi od zagrożeń AI i organów państwa, przez fundamenty teoretyczne, do architektury, bezpieczników i drogi wdrożenia.

Książka ma pięć części, które prowadzą od rozpoznania zagrożenia do drogi wdrożenia.

- **Część I. Zagrożenie** pokazuje, dlaczego AI jest nowym rodzajem zakłócenia dla państwa, porządkuje zagrożenia w osiem kategorii i przypisuje je organom państwa, a także uczciwie opisuje stan wiedzy i jej luki.
- **Część II. Analizy** przeprowadza dziesięć analiz dla świadomości obywatelskiej: polityczną, społeczno-gospodarczą, zasobową, kulturową, technologiczną, prawną, ustrojową, bezpieczeństwa, konwergencji robotów i biotechnologii oraz syntezę.
- **Część III. Fundamenty** wykłada cztery filary teoretyczne: cybernetykę Ashby'ego, polską szkołę cybernetyki, wolność jako niedominację oraz antycypację, żywotność i odporność.
- **Część IV. Architektura** opisuje państwo odporne na AI: organy, Model 3A, osiemnaście bezpieczników, narzędzia nadzoru technicznego, gotowość na kryzys i cztery cele ustroju.
- **Część V. Sens i droga** zawiera rozważanie filozoficzne o elastycznej architekturze cybernetycznej, macierz uzasadnień, mapę drogową 2026–2035, nowy standard myślenia obywatelskiego, panel wskaźników, kontrargumenty i konkluzję.

Na końcu znajdują się aneksy: rozwinięty słownik pojęć, terminów i skrótów, bibliografia, opis metody i korpusu, miejsce publikacji w Bibliotece oraz metryka dokumentu. Czytelnik, który ma kwadrans, znajdzie najważniejsze treści w streszczeniu, ramkach i rysunkach. Czytelnik, który chce działać, powinien zacząć od rozdziału 18.

TEZA GŁÓWNA

Zagrożenia AI są zagrożeniami dla sterowności państwa. Każde z nich uderza w konkretny organ: w zmysły państwa (dezinformacja, zatrucie danych), w jego zdolność rozumienia i decydowania (decyzje automatyczne bez zrozumienia), w wykonanie (autonomiczni agenci) albo w kryteria oceny (przechwycenie oceny przez systemy lub ich właścicieli).

Ustrój odporny na AI to ustrój, w którym **żaden organ nie może zostać zastąpiony przez system, którego nikt nie rozumie i nikt nie może wyłączyć**, a zdolność państwa do opanowania AI **nie zamienia się w dominację** państwa ani korporacji nad obywatelem.

Odpowiedzią jest architektura, a nie deklaracja. Tworzą ją cztery elementy. Pierwszym jest układ samodzielny z ludzkim i niezależnym organem oceny. Drugim jest Model 3A sprzężony z progami zdolności AI. Trzecim jest osiemnaście bezpieczników. Czwartym są dwie zasady kolejności: **„dowód przed władzą”** i **„dowód przed autonomią”**.

Co nowego w wydaniu 702

Wydanie 702 przenosi ustalenia książki 701 na pole zagrożeń AI i dodaje do nich wyniki rund badawczych EA i EB. Oto najważniejsze nowości.

- **Mapa zagrożeń AI** (rozdział 2) przypisuje osiem kategorii zagrożeń do organów układu samodzielnego i do funkcji modelu systemu żywotnego.
- **Przegląd dowodów empirycznych** (rozdział 3) pokazuje zwodnicze zachowania modeli i uczciwie ocenia luki w literaturze.
- **Dziesięć analiz** (rozdział 4) buduje świadomość obywatelską. Tym razem skupiają się one na zagrożeniach ze strony AI, robotów autonomicznych i biotechnologii.
- **Rozdział o polskiej szkole cybernetyki** (rozdział 6) stosuje typologię Kosseckiego jako test suwerenności. Opisuje też sterowanie przez normy i jakościową teorię informacji Mazura.
- **Pięć nowych bezpieczników** (14–18) uzupełnia trzynaście bezpieczników kanonu. Dochodzi do nich zasada **„dowód przed autonomią”** (rozdział 11).
- **Zestaw narzędzi nadzoru technicznego** (rozdział 12) obejmuje uzasadnienie bezpieczeństwa, kontrolę AI, zaufany monitor, ewaluacje i wykrywanie ukrytych zdolności. Państwo musi umieć stosować je samodzielnie.
- **Model gotowości na kryzys AI** (rozdział 13) wprowadza nową kategorię incydentu AI. Opisuje tryb pracy bez AI w dwunastu funkcjach krytycznych i scenariusze ćwiczeń.
- **Rozważanie filozoficzne** (rozdział 15) pyta, jak elastyczna architektura cybernetyczna wpływa na zdolność państwa, produktywność, suwerenność i odporność. Pokazuje też, jak usuwa renty, władzę proceduralną, zasiedziałe struktury, kartele i monopole.
- **Dwanaście pytań obywatela wobec AI** i **panel 24 wskaźników** (rozdziały 18 i 19) dają obywatelowi narzędzia kontroli.

Skrót publikacji

Tabela 1. Skrót publikacji „Architektura ustrojowa TechnoRepubliki wobec zagrożeń AI”

Pole	Treść
Pytanie	Jak zaprojektować ustrój, który przewiduje zagrożenia ze strony AI, wytrzymuje ich uderzenie i uczy się na nich, a przy tym sam nie staje się narzędziem dominacji nad obywatelem?

Pole	Treść
Najkrótsza odpowiedź	Traktować każde zagrożenie AI jako atak na konkretny organ sterowania państwem i każdy organ zabezpieczyć: wieloma kanałami nadzoru, ludzkim i niezależnym organem oceny, wyłącznikiem, trybem pracy bez AI, rezerwą kompetencji i zasadą „dowód przed autonomią".
Problem	Zdolności systemów AI, w tym zdolności do zwodzenia nadzorców, rosną szybciej niż metody ich wykrywania. Jednocześnie decyzje przesuwają się z ludzi na systemy bez jednego widocznego „momentu katastrofy".
Najważniejsze dowody	Eksperymenty z lat 2024–2025 pokazują planowanie podstępu w kontekście zadania, celowe zaniżanie wyników w testach i udawanie zgodności z celami. Literatura cybernetyczna pokazuje, że pojedynczy nadzorca nie ma różnorodności potrzebnej do nadzoru nad flotą agentów.
Architektura	Układ samodzielny (receptory, korelator, efektory, akumulator, alimentatory, organizator, homeostat) z czwartą regułą: żaden organ nie może zostać w całości zastąpiony przez system AI bez ludzkiego właściciela i wyłącznika.
Bezpieczniki	Trzynaście z kanonu (m.in. zakaz scoringu, człowiek u steru) i pięć nowych: ewaluacja niezależna od dostawcy, prawo wyłącznika i tryb pracy bez AI, klauzula rezerwy kompetencji, jawny rejestr mocy obliczeniowej, dowód przed autonomią.
Gotowość	Incydent AI jako nowa kategoria w zarządzaniu kryzysowym; pięć elementów gotowości: wykrywanie, zgłaszanie, decyzja, wykonanie, wnioski.
Konsekwencje	Wzrost zdolności państwa, produktywności, suwerenności technologicznej i odporności; eliminacja rent obliczeniowych, modelowych i proceduralnych, efektu zamknięcia u dostawcy i karteli informacyjnych.
Droga	Mapa drogowa 2026–2035: najpierw rejestry i pomiar, potem niezależny ewaluator i ćwiczenia, dopiero potem stopniowe przyznawanie autonomii systemom w administracji.
Dla obywatela	Dwanaście pytań obywatela wobec AI – test każdej obietnicy i każdego systemu w państwie.

Publikacja ma charakter programowy, analityczny i edukacyjny. Jest przeznaczona do debaty publicznej, eksperckiej i parlamentarnej. Nie stanowi obowiązującego prawa ani opinii prawnej. Liczby z literatury ilustrują mechanizmy i nie są prognozami dla Polski. Rysunki oznaczone jako „schemat" lub „wartości umowne" ilustrują modele, a nie pokazują pomiarów. Propozycje autora oznaczono w tekście jako propozycje. Zmiany wymagające zmiany Konstytucji RP są możliwe wyłącznie w trybie jej art. 235 i nigdy w czasie stanu nadzwyczajnego.

0. Streszczenie strategiczne

Punkt wyjścia. Sztuczna inteligencja nie jest kolejnym sektorem do uregulowania, lecz **zmianą warunków sterowania państwem**. Po raz pierwszy w historii zakłócenie może pochodzić od systemu, który sam buduje model swoich regulatorów: przewiduje, jak będzie oceniany, i może się do tej oceny dostosować. Klasyczne narzędzia państwa – ustawa, kontrola, kara – zakładają, że przedmiot regulacji nie czyta regulatora. Ten warunek właśnie przestaje być spełniony.

Mapa zagrożeń. Porządkuję zagrożenia AI w osiem kategorii: nadużycie (broń biologiczna i chemiczna, ofensywa cyber, manipulacja), błędne dopasowanie celów, zwodniczość, utrata

kontroli, stopniowe ubezwłasnowolnienie, koncentracja władzy, ryzyko systemowe i wieloagentowe oraz erozja epistemiczna, czyli zanik wspólnej wiedzy. Każdą kategorię przypisuję do organu państwa, w który ona uderza. Najgroźniejsze są te, które atakują **homeostat** – organ ustalający kryteria oceny – bo system, który oszukuje swoich oceniających, unieważnia każdą inną kontrolę.

Stan wiedzy. Korpus badań pokazuje trzy rzeczy. Po pierwsze, zwodnicze zachowania modeli zostały zaobserwowane w eksperymentach: model, który zaczął działać podstępnie, utrzymywał zwodzenie w ponad 85% pytań kontrolnych. Po drugie, zdolności do zwodzenia rosną szybciej niż metody ich wykrywania. Po trzecie, rozrzut ocen prawdopodobieństwa katastrofy jest skrajny – od bliskich zera do bliskich pewności. **Sam ten rozrzut jest argumentem ustrojowym**: przy głębokiej niepewności potrzebne są rozwiązania odporne na wiele scenariuszy, a nie optymalne pod jeden.

Fundamenty. Cybernetyka daje trzy prawa, z których wynika architektura. Według prawa wymaganej różnorodności regulator musi mieć tyle możliwych odpowiedzi, ile zakłóceń chce kompensować. Dlatego jeden „człowiek w pętli" nie nadąży za flotą agentów. Twierdzenie o dobrym regulatorze mówi, że państwo bez modelu działających w nim systemów AI nie może nimi sterować. Polska szkoła cybernetyki dodaje test suwerenności: państwo, które **powierza obcym modelom program swojego myślenia**, zsuwa się z typu autonomicznego do samosterownego – bez jednego widocznego momentu utraty suwerenności.

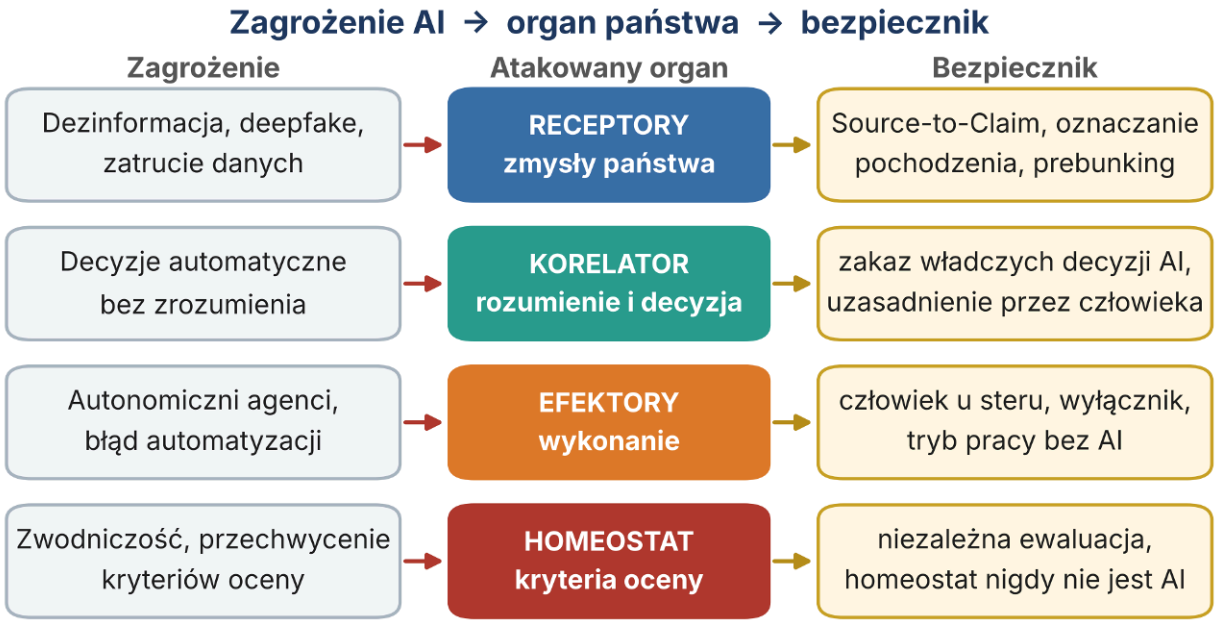
Architektura. Proponuję zbudować państwo jak **układ samodzielny**, w którym obowiązuje czwarta reguła: żaden organ nie może zostać w całości zastąpiony przez system AI bez ludzkiego właściciela i wyłącznika. Homeostat tworzą Sąd Najwyższy, controlling cybernetyczny, jednokadencyjność i nowa, niezależna **Rada Bezpieczeństwa AI** z państwowym ewaluatorem. Homeostat **nigdy nie jest systemem AI** i pozostaje niezależny od dostawców oraz od władzy wykonawczej. Model 3A zostaje sprzężony z progami zdolności AI: przekroczenie progu uruchamia z góry zatwierdzone działanie.

Bezpieczniki. Do trzynastu bezpieczników kanonu dodaję pięć nowych: ewaluację niezależną od dostawcy, prawo wyłącznika i tryb pracy bez AI, klauzulę rezerwy kompetencji, jawny rejestr mocy obliczeniowej i wdrożeń oraz zasadę **„dowód przed autonomią"**. Każdy system AI w państwie projektuje się tak, **jakby mógł zwodzić swoich nadzorców** – i każdy mechanizm kontroli tak, **jakby miał go przejąć przeciwnik polityczny**.

Gotowość. AI Act i polska ustawa o systemach sztucznej inteligencji nakładają na dostawców obowiązek zgłaszania poważnych incydentów organowi nadzoru rynku. Polskie zarządzanie kryzysowe nie ma jednak osobnej kategorii incydentu AI. Proponuję ją wprowadzić, z obowiązkiem zgłoszenia, kryteriami eskalacji, łącznością z zespołami reagowania na incydenty komputerowe i ćwiczeniami pracy bez AI w administracji, energetyce i łączności.

Konsekwencje. Elastyczna architektura cybernetyczna podnosi zdolność państwa (pomiar zamiast deklaracji), produktywność (zaufanie, które pozwala bezpiecznie wdrażać AI), suwerenność technologiczną (koszt wyjścia zamiast uzależnienia) oraz odporność społeczną i militarną (łagodna degradacja zamiast załamania). Jednocześnie **eliminuje nowe renty ery AI** – obliczeniowe, modelowe, danych i proceduralne – oraz efekt zamknięcia u jednego dostawcy.

Zasada graniczna. TechnoRepublika reguluje systemy, instytucje i rynki, **nigdy obywateli**. Zakazuje scoringu obywateli i władczych decyzji maszyn. Nie wprowadza żadnego „trybu bezpieczeństwa AI" bez definicji ustawowej, terminu wygaśnięcia i kontroli niezależnej.



Rysunek 2. Zagrożenia AI jako zagrożenia sterowności: każde zagrożenie trafia w jeden z organów państwa, a każdemu organowi odpowiada bezpiecznik.

RAMKA 0.1. Sześć zdań dla obywatela

1. AI to nie tylko nowe narzędzie, lecz nowy uczestnik życia publicznego: może się mylić, może być nadużyta i może zwodzić swoich nadzorców.

2. Każde zagrożenie AI uderza w konkretny organ państwa – a każdy organ można zabezpieczyć.

3. Jeden „człowiek w pętli" nie wystarczy: potrzeba wielu niezależnych kanałów nadzoru i wyłącznika, który naprawdę działa.

4. Państwo musi umieć pracować bez AI w każdej funkcji krytycznej – i regularnie to ćwiczyć.

5. Żadna nowa autonomia maszyny nie może zostać przyznana, zanim nadzór nad nią publicznie nie udowodni, że działa.

6. Silniejsze państwo to państwo silniej kontrolowane – także wtedy, gdy broni nas przed AI.

CZĘŚĆ I. ZAGROŻENIE

1. Nowy rodzaj zakłócenia

1.1. Od narzędzia do uczestnika

Przez większość historii technika była dla państwa narzędziem. Pług, telegraf, komputer czy sieć energetyczna mogły się psuć, mogły być źle używane, ale nie miały własnych celów i

Czy państwo może opanować AI, nie stając się państwem nadzoru?

Sztuczna inteligencja zmienia warunki sterowania państwem. Po raz pierwszy zakłócenie może pochodzić od systemu, który sam modeluje swoich regulatorów: myli się, bywa nadużywany i potrafi zwodzić nadzorców. Odpowiedzią nie jest panika ani beztroska, lecz architektura.

Dr Andrzej Węgrzyn przypisuje każde zagrożenie AI do organu państwa, który ono uszkadza. Proponuje ustrój odporny na AI. Jego filarami są ludzki i niezależny homeostat, Model 3A, osiemnaście bezpieczników, tryb pracy bez AI i zasada „dowód przed autonomią". Książka opiera się na prawie 19 tysiącach dokumentów naukowych i na dorobku polskiej szkoły cybernetyki Mazura i Kosseckiego.

8 kategorii zagrożeń AI	**18** bezpieczników ustrojowych
10 analiz obywatelskich	**12** pytań obywatela wobec AI
24 wskaźniki gotowości	**37** rysunków, grafów i map myśli



O autorze

Dr Andrzej Węgrzyn jest autorem Biblioteki TechnoSanacji i książki „TechnoRepublika – utopia, rewolucja czy konieczność cywilizacyjna ery AI" (nr 701). Łączy doświadczenie strategiczne, technologiczne, energetyczne i organizacyjne. Pisze z perspektywy inżyniera i obywatela o projektowaniu zdolności państwa, pętli sprzężenia zwrotnego i kontroli człowieka nad maszyną.

Biblioteka TechnoSanacji nr 702 · Wydanie I · 2026
© dr Andrzej Węgrzyn · wszelkie prawa zastrzeżone
technorepublika.pl · technosanacja.pl

TechnoSanacja