**Nowoczesne Systemy Zarządzania**
Zeszyt 18 (2023), nr 3 (lipiec-wrzesień)
ISSN 1896-9380, s. 31-44
DOI: 10.37055/nsz/183866

Instytut Organizacji i Zarządzania
Wydział Bezpieczeństwa, Logistyki i Zarządzania
Wojskowa Akademia Techniczna
w Warszawie

**Modern Management Systems**
Volume 18 (2023), No. 3 (July-September)
ISSN 1896-9380, pp. 31-44
DOI: 10.37055/nsz/183866

Institute of Organization and Management
Faculty of Security, Logistics and Management
Military University of Technology
in Warsaw

# Dylematy etyczne związane z ewolucją robotów

# Ethical dilemmas related to the robot evolution

**Małgorzata Maternowska**
Akademia Górniczo-Hutnicza w Krakowie
mmaterno@zarz.agh.edu.pl; ORCID: 0000-0003-0971-7895

**Abstrakt.** Rozwój technik obliczeniowych, robotyki, druku 3D i technologii materiałowych umożliwia tworzenie zaawansowanych systemów robotów, które mogą autonomicznie rozmnażać się i ewoluować. Powstająca technologia ewolucji robotów rzuca wyzwanie istniejącej etyce sztucznej inteligencji, ponieważ wrodzona adaptacyjność, stochastyczność i złożoność systemów ewolucyjnych stwarzają liczne zagrożenia. Trudno jest lekceważyć możliwe implikacje dwóch kluczowych funkcjonalności ewoluujących robotów: samoreplikacji i losowej zmiany formy oraz zachowania robota. Samoreplikacja umożliwia robotom rozmnażanie się bez interwencji człowieka. Mutacje lub losowe zmiany ewolucyjne mogą prowadzić do niepożądanych i szkodliwych zachowań robotów, zagrażając ludzkim interesom. Za każdym razem, gdy powstaje technologia, która nie jest bezpośrednio kontrolowana przez człowieka, i gdy proces ten jest nieprzewidywalny, rodzą się pytania o ryzyko i odpowiedzialność. W artykule porusza się kwestie możliwego ryzyka powstania szkód i odpowiedzialności w powiązaniu z kluczowym problemem kontroli ludzkiej nad procesem ewolucji. Zagadnienie odpowiedzialności za sztuczną inteligencję uznano za szczególnie istotne zarówno z etycznego, jak i prawnego punktu widzenia. Generalnie odpowiedzialność dotyczy pełnego spektrum zdarzeń *ex post* (kto zawinił, jaki był powód) oraz *ex ante* (jakie działania należy podjąć, by zmniejszyć ryzyko, czyli co jest zaniedbaniem i kto je popełnia) i z reguły spoczywa na ludziach. Problemem jest jednak istnienie luk w zakresie odpowiedzialności za powstałe szkody czy zdarzenia niepożądane, w których uczestniczą ludzie i systemy sztucznej inteligencji, jakich nie da się wypełnić tradycyjnymi koncepcjami przypisania odpowiedzialności. W artykule wskazano na proponowany w literaturze przedmiotu sposób kompleksowego rozwiązania problemu luk w zakresie odpowiedzialności za sztuczną inteligencję, oparty na idei projektowania systemów socjotechnicznych umożliwiających znaczącą kontrolę człowieka, czyli systemów dostosowanych do ludzkich intencji i możliwości.
Celem artykułu jest wskazanie na pilną potrzebę ciągłego poszukiwania odpowiedzi na pytanie: w jaki sposób można odpowiedzialnie kontrolować ewolucję robotów?
**Słowa kluczowe:** robotyka ewolucyjna, etyka, prawo cyberprzestrzeni, znacząca kontrola człowieka, luki w odpowiedzialności

**Abstract.** Development of computational techniques, robotics, 3D printing and materials technologies enables the creation of advanced robotic systems that can autonomously reproduce and evolve. Emerging robotic evolution technology challenges existing AI ethics. The inherent adaptability, stochasticity and complexity of evolutionary systems pose numerous threats. It is difficult to underestimate the possible implications of two key functionalities of evolving robots: self-replication and random change of robot form and behavior. Self-replication allows robots to reproduce without human intervention. Mutations or random evolutionary changes can lead to undesirable and harmful behavior of robots, threatening human interests. Whenever a technology is developed that is not under direct human control and the process is unpredictable, questions of risk and liability arise. The article addresses both possible risk of harm and liability in connection with the key issue of human control over the process of evolution. The issue of responsibility for artificial intelligence was considered to be particularly important from both an ethical and legal point of view. In general, responsibility covers the full spectrum of events ex post (who is at fault, what was the reason), and ex ante (what actions should be taken to reduce the risk, i.e. what is negligence and who commits it), and usually rests with people. The problem, however, is the existence of gaps in the scope of liability for damages or adverse events involving people and artificial intelligence systems that cannot be filled with traditional concepts of assigning responsibility. The article indicates a method of comprehensively solving the problem of gaps in liability for artificial intelligence, proposed in the literature, based on the idea of designing social engineering systems that enable "meaningful human control", i.e. systems adapted to human intentions and capabilities.
The aim of the article is to indicate the urgent need for a continuous search for an answer to the question of how to responsibly control the evolution of robots.
**Keywords:** evolutionary robotics, ethics, cyber low, meaningful human control, responsibility gaps

## Wstęp

Można powiedzieć, że idea ewolucji robotów ma już sto lat. Słynna sztuka Karela Čapka[1], który jako pierwszy, zupełnie niechcący, użył słowa „robot", została opublikowana w 1920 r. Pod koniec utworu roboty są na skraju wyginięcia, a jeden z ludzi, Alquist, radzi im: „Jeśli chcesz żyć, musisz się rozmnażać jak zwierzęta".

W roku 1920 był to fantastyczny pomysł, ale nierealny. We współczesnym świecie, z szybko rozwijającą się sztuczną inteligencją i robotyką, to wciąż fantastyczny pomysł, ale już nie niemożliwy. Pod koniec XX wieku przeniesiono zasady ewolucji biologicznej do sfery technologii wdrażanych w symulacjach komputerowych. Algorytmy ewolucyjne okazały się zdolne do rozwiązywania trudnych problemów z różnych dziedzin naukowych i technicznych, oferując przewagę nad tradycyjnymi metodami optymalizacji i projektowania (Ashlock, 2006; de Jong, 2006; Eiben, Smith, 2003). Algorytmy ewolucyjne zostały również zastosowane do opracowania sprzętu (ang. *hardware*) i oprogramowania (ang. *software*) robotów autonomicznych, co zaowocowało powstaniem nowej dziedziny o nazwie robotyka ewolucyjna (Nolfi, Floreano, 2000; Bongard, 2013; Vargas, di Paolo, Harvey et al., 2014; Doncieux, Bredeche, Mouret, Eiben, 2015).

Do tej pory prace nad robotyką ewolucyjną prowadzono głównie w ramach symulacji komputerowych. Nieliczne doniesienia naukowe zawierały informacje

[1] R.U.R. – Rossum's Universal Robots (Čapek, 1920).

o samoreprodukujących się maszynach fizycznych (Lipson, Pollack, 2000; Kriegman, Blackiston, Levin, Bongard, 2020). Nie były to jednak przykłady rozwiązań dotyczących *stricte* ewolucji ich systemów; w reprodukcji powstawały identyczne, niezmieniające się klony.

Sytuacja ta jednak szybko się zmienia. Po pierwszym dużym osiągnięciu, jakim w dziedzinie robotyki i sztucznej inteligencji (ang. *Artificial Itelligence*, AI) było przejście od *wetware*[2] do *software* w XX wieku, ewolucja jest u progu drugiego znaczącego etapu. Tym razem dotyczy to przejścia od *software* do *hardware*[3] (Ellery, 2020).

Nauka i technologia związane z ewolucją robotów wynoszą obawy dotyczące sztucznej inteligencji i robotyki na nowy poziom, przede wszystkim dzięki zjawisku, które zwykło nazywać się *second order engineering* lub *second order design* (Eiben, Ellers, Meynen, Nyholm, 2021). Obecna praktyka to *first order system*: sztuczna inteligencja i roboty są opracowywane i skonstruowane bezpośrednio przez ludzi. Ewolucyjna technologia robotów radykalnie zmienia ten obraz; zamiast konstruować robota do realizacji określonego zadania, tworzy się system ewolucyjny, który umożliwia powstanie robota. Kwestie etyczne, moralne i kwestie bezpieczeństwa powinny zatem zostać przekształcone w zasady projektowania i wskazówki metodologiczne dla ludzi.

Nowe wyzwania etyczne związane z ewolucją robotów są zakorzenione we wrodzonej nieprzewidywalności procesu ewolucyjnego. Ewolucja przebiega w wyniku generowania dziedzicznej zmienności (rekombinacji i mutacji) w połączeniu z selekcją, która faworyzuje bardziej udane formy kosztem dużej liczby nietrafionych. Rozwój robotów przez zautomatyzowaną (re)produkcję może zatem doprowadzić do powstania znacznej liczby dowolnych form robotów, co zwiększa prawdopodobieństwo niezamierzonego tworzenia robotów o szkodliwych zachowaniach. Co więcej, kluczowe zmiany ewolucyjne często zachodzą w postaci innowacji, które wynikają z przeorganizowania istniejących cech dla nowych funkcji. Dokonująca się w taki sposób ewolucja jest wysoce nieprzewidywalna zarówno pod względem kierunku, jak i wielkości, zwiększając tym samym prawdopodobieństwo, że rozwijające się roboty będą miały nieoczekiwane możliwości. Oprócz tego mutacja lub losowe zmiany ewolucyjne w projekcie mogą prowadzić do niepożądanych zachowań robotów, działając na szkodę ludzkich interesów. Przed opracowaniem nowej technologii z tak potencjalnie dużymi konsekwencjami powinno się zatem określić poziom akceptowalności możliwych konsekwencji oraz sposoby przewidywania niepożądanych skutków.

---

[2] Technologia komputerowa, w przypadku której ludzki mózg używany jest jako model sztucznych systemów opartych na procesach biochemicznych.

[3] W roku 2019 inżynierowie z Uniwersytetu Cornella na podstawie technologii, którą nazwali DASH (*DNA-based Assembly* i *Synthesis of Hierarchical*), stworzyli syntetyczny materiał wykazujący trzy kluczowe cechy życia: metabolizm, zdolność samomontażu i samoorganizację. Zbudowane z niego maszyny zachowują się zupełnie tak jak żywe organizmy. Są zdolne do poruszania się i metabolizmu, pochłaniają energię, rozwijają się i rozpadają, a także ewoluują (podano za: Usidus, 2023).

Za każdym razem gdy powstaje technologia, która nie jest bezpośrednio kontrolowana przez człowieka (technologie bez „kierownicy"), i gdy proces ten jest nieprzewidywalny, rodzą się pytania o ryzyko i odpowiedzialność (Santoni de Sio, van den Hoven, 2018; Nyholm, 2020). Czy korzyści z użytkowania nowej technologii przewyższają jej możliwe negatywne skutki? Jeśli występują negatywne skutki, jak możemy je minimalizować i kontrolować? Wreszcie, co jest ważne, gdy sprawy wymykają się spod kontroli, kto jest za to odpowiedzialny (Maternowska, 2022)? Odpowiedź na te pytania wymaga nie tylko rozwiązań z zakresu techniki i technologii, lecz także rozstrzygnięć rodzących się problemów etycznych, które dotyczą podejmowania działań zapobiegających ewentualnym szkodom. Można argumentować, że takie obawy są na obecnym etapie rozwoju niepotrzebne, jednak jeśli zacznie się myśleć o możliwych negatywnych skutkach w momencie, gdy te zaistnieją, to najprawdopodobniej będzie już za późno.

## Ryzyko

Choć pomysł eksplozji sztucznej inteligencji z udziałem samoreplikujących się, superinteligentnych maszyn wydaje się być może zbyt futurystyczny, pojawiają się opinie, zarówno w środowisku akademickim, jak i poza nim, które traktują go bardzo poważnie. Wśród tych, którzy wyrażają/wyrażali swoje obawy, są/byli filozofowie, tacy jak Nick Bostrom[4] i Toby Ord[5], czy też wybitne osobistości, takie jak Elon Musk i nieżyjący już Stephen Hawking.

Już teraz trudno zlekceważyć możliwe implikacje dwóch kluczowych funkcjonalności ewoluujących robotów: samoreplikacji i losowej zmiany formy i zachowania robota. Samoreplikacja umożliwia robotom rozmnażanie się bez interwencji człowieka. Mutacje lub losowe zmiany ewolucyjne mogą prowadzić do niepożądanych i szkodliwych zachowań robotów, zagrażając ludzkim interesom.

Kilka dziedzin nauki doświadczyło podobnego dylematu w zakresie bezpieczeństwa podczas rozwoju nowych technologii i późniejszych eksperymentów. Biomedyczne dylematy etyczne są zazwyczaj rozstrzygane na podstawie zasad, takich jak autonomia, nieszkodzenie (unikanie krzywdy), działanie „dla dobra" i sprawiedliwość (za: Eiben, Ellers, Meynen, Nyholm, 2021). W kontekście eksperymentów technologicznych dodano pojęcie odpowiedzialności (van de Poel, 2016), a konkretnie, w dziedzinie sztucznej inteligencji, pojęcie wyjaśnialności (ang. *explicability*), co

[4] Szwedzki filozof, profesor Uniwersytetu Oksfordzkiego, autor licznych prac dotyczących transhumanizmu (idei zakładającej wykorzystanie osiągnięć nauki i techniki w celu przezwyciężenia ludzkich ograniczeń). Magazyn „Foreign Policy" umieścił go na liście 100 czołowych myślicieli świata.

[5] Australijski filozof i etyk, związany z efektywnym altruizmem. Jest pracownikiem naukowym Instytutu Przyszłości Ludzkości w Oxfordzie. W swojej pracy badawczej skupia się na kwestiach związanych z ryzykiem egzystencjalnym.

oznacza, że gdy algorytmy oparte na sztucznej inteligencji są wykorzystywane do podejmowania moralnie wrażliwych decyzji, ludzie powinni być w stanie uzyskać rzeczywiste, bezpośrednie i jasne wyjaśnienie procesu decyzyjnego lub uzasadnienie decyzji wynikającej z algorytmu (Floridi, Cowls, Beltrametti et al., 2018).

W robotyce ewolucyjnej wszystkie te zasady mają istotne znaczenie. Ryzyko szkody i kwestia odpowiedzialności musi być jednak rozważona szczegółowo, co z kolei jest ściśle związane z kluczową kwestią kontroli i potencjalną jej utratą. Zwykle uważa się, że aby konkretna istota ludzka lub grupa ludzi była odpowiedzialna za jakiś proces lub jego wynik, musi mieć możliwość kontrolowania tego procesu lub jego wyniku. Co więcej, utratę kontroli można uznać za formę szkody, ponieważ zazwyczaj postrzega się ją jako podważanie ludzkiej autonomii oraz zagrożenie dla innych wartości (takich jak na przykład dobrostan), które w pewnym stopniu zależą od naszej zdolności do kontrolowania tego, co się wokół nas dzieje.

Największy problem w przewidywaniu możliwych zagrożeń w przypadku ewoluujących robotów polega na tym, że ma się do czynienia z systemem, który z natury ciągle się zmienia. Ryzyko szkody musi zatem być oceniane pod kątem potencjalnej trajektorii przyszłych zmian i kierunków rozwoju procesu ewolucyjnego.

Autorzy publikacji *Robot Evolution: Ethical Concerns* (Eiben, Ellers, Meynen, Nyholm, 2021) wyróżnili **trzy kluczowe rodzaje ryzyka**:

- **ryzyko związane z reprodukcją:** roboty mogą ewoluować, szybko się reprodukując, co skutkuje niekontrolowanym wzrostem ich populacji. Jeśli populacja robotów stanie się zbyt duża, zasoby, takie jak przestrzeń, energia, surowce, powietrze lub woda, mogą zostać (lokalnie) wyczerpane. Efekt ten można porównać do plagi szarańczy;
- **ryzyko nieprzystosowania:** ewolucja robotów stworzonych do realizacji określonego zadania może prowadzić do sytuacji, gdy pewne cechy lub zachowanie robota, korzystne ze względu na realizowane przez niego zadanie, mogą być szkodliwe dla ludzi. W ekstremalnych przypadkach roboty mogą szkodzić ludziom, np. gdy ludzie utrudniają im wykonywanie zadania. Ten rodzaj ryzyka może ewoluować, ponieważ selekcja stosowanych przez robota rozwiązań jest „ślepa", czyli zawsze zwyciężają najskuteczniejsze rozwiązania bez względu na ewentualne konsekwencje dla otoczenia;
- **ryzyko dominacji:** aby stać się dominującym „gatunkiem", roboty mogą ewoluować w związku z określoną cechą związaną z ich funkcjonalnością (nie chodzi w tym przypadku o efekt selekcji). Może się tak zdarzyć, że staną się lepsze od ludzi pod względem intelektualnym, fizycznym lub emocjonalnym (będąc stabilnymi i konsekwentnymi)[6]. W rezultacie mogą

[6] „To jest tylko kwestia kilku, kilkunastu, najdłużej kilkudziesięciu lat, kiedy będziemy pod każdym względem daleko w tyle za sztuczną inteligencją" – tak przewiduje prof. Andrzej Dragan, fizyk kwantowy, w wywiadzie dla portalu Wirtualna Polska (Wyżga, Siedzik, 2023).

decydować, pośrednio lub bezpośrednio organizując życie ludziom i tym samym zmniejszając ich autonomię. Efekt ten można porównać do relacji rodzic – dziecko, w której rodzic lepiej rozumie i przewiduje sytuacje, a zatem ogranicza zasięg przestrzenny i czynności dziecka.

## Luki w odpowiedzialności

Podstawową cechą sztucznej inteligencji jest rozwiązywanie złożonych problemów na podstawie zgromadzonych zasobów danych, wyciąganie wniosków oraz samodzielne (autonomiczne) podejmowanie decyzji dla osiągnięcia z góry założonego przez programistę czy dysponenta celu (Maternowska, 2019). Niezależny od twórcy rozwój sztucznej inteligencji-robotów w połączeniu z autonomicznością ich działań może sprawiać, że decyzje podejmowane przez takie systemy będą trudne do przewidzenia oraz uzasadnienia. Pojawia się zatem pytanie o skutki tych decyzji i towarzyszące im kwestie odpowiedzialności.

W sytuacji, gdy w grę wchodzi opisane wcześniej ryzyko utraty możliwości kontroli nad procesem ewolucji, określenie odpowiedzialności jest ważne, zarówno z etycznego, jak i prawnego punktu widzenia. Odpowiedzialność ta dotyczy pełnego spektrum zdarzeń *ex post* (kto zawinił, jaki był powód) oraz *ex ante* (jakie działania należy podjąć, by zmniejszyć ryzyko, czyli co jest zaniedbaniem i kto je popełnia) i z reguły spoczywa na ludziach.

Tradycyjnie producent/operator maszyny ponosi moralną i prawną odpowiedzialność za skutki jej eksploatacji. Autonomiczne, uczące się maszyny, oparte na sieciach neuronowych, algorytmach genetycznych i architekturach agentów, stwarzają sytuację, w której producent/operator maszyny w zasadzie nie jest już w stanie przewidzieć przyszłego zachowania maszyny, a tym samym nie może być pociągnięty do odpowiedzialności moralnej. Społeczeństwo musi zdecydować, czy nie używać tego rodzaju maszyn (co nie jest realistyczną opcją), czy stanąć w obliczu luki odpowiedzialności, której nie da się wypełnić tradycyjnymi koncepcjami przypisania odpowiedzialności (Matthias, 2004; Nyholm, 2020).

Pojęcie „luka odpowiedzialności" zostało wprowadzone do debaty filozoficznej przez Andreasa Matthiasa (2004), aby wskazać na obawę (ryzyko), że uczenie się automatów może utrudnić lub uniemożliwić przypisywanie ludziom winy za zdarzenia niepożądane. Obecnie przyjmuje się, iż źródłem tego ryzyka jest nie tyle uczenie maszynowe, ile ogólnie nieprzejrzystość, złożoność i nieprzewidywalność systemów sztucznej inteligencji (Amoroso, Tamburrini, 2019).

Filippo Santoni de Sio i Giulio Mecacci (2021) wyodrębnili cztery rodzaje luk odpowiedzialności, które w przypadku sztucznej inteligencji mają największe znaczenie (zob. tabelę 1).

Tabela 1. Cztery rodzaje luk odpowiedzialności

| | Rodzaj odpowiedzialności | Definicja | Luki w odpowiedzialności związane ze sztuczną inteligencją |
|---|---|---|---|
| **Odpowiedzialność bierna** | Odpowiedzialność na zasadzie winy[7] | Opiera się na etycznym założeniu, że ten, kto swoim zawinionym czynem (zamiar, wiedza lub kontrola) wyrządził komuś szkodę, powinien ponosić konsekwencję swego zachowania | Sztuczna inteligencja utrudnia przewidywanie i kontrolę, a tym samym określenie uzasadnionych przyczyn nagannego postępowania, np. możliwe do uniknięcia zderzenie drogowe z udziałem automatu, system napędowy, którego nikt nie byłby w stanie samodzielnie przewidzieć lub mu zapobiec |
| | Odpowiedzialność moralna | Obowiązek wyjaśnienia swoich racji i działania wobec innych (w pewnych okolicznościach) | Sztuczna inteligencja sprawia, że procesy stają się niewytłumaczalne dla korzystających z niej osób, np. lekarz nieumiejący wytłumaczyć pacjentowi podstaw swojej diagnozy postawionej z zastosowaniem sztucznej inteligencji |
| | Odpowiedzialność publiczna | Obowiązek wyjaśniania swoich działań opinii publicznej | Sztuczna inteligencja przenosi uprawnienia dyskrecjonalne na ekspertów IT i analityków danych. Często są to prywatne firmy, których działalność jest trudniejsza do publicznego zbadania, np. wykorzystywanie (prywatnych) systemów sztucznej inteligencji do wsparcia podejmowania decyzji na szczeblu rządowym |
| **Odpowiedzialność czynna** | Odpowiedzialność na zasadzie słuszności | Obowiązek promowania i osiągania wspólnych celów i wartości | Podmioty zaangażowane w projektowanie lub wykorzystywanie sztucznej inteligencji nie są wystarczająco świadome własnej odpowiedzialności za zapobieganie szkodom wynikającym ze sztucznej inteligencji lub nie mają możliwości (motywacji ) wypełnienia tego obowiązku, m.in. dotyczy to inżynierów lub menedżerów patrzących tylko na techniczne korzyści płynące z zastosowania AI |

Źródło: opracowanie własne na podstawie: Santoni de Sio, Mecacci, 2021

[7] Należy wspomnieć, iż luka związana z odpowiedzialnością na zasadzie winy nie powstała przez pojawienie się „uczących się automatów" wraz z ich wrodzoną nieprzewidywalnością, jak to zostało sformułowane przez A. Matthiasa (2004). W rzeczywistości inne inteligentne, autonomiczne byty „bez duszy do obwiniania i bez ciała do kopania", takie jak biurokracje i korporacje, mogą same w sobie generować luki w zakresie winy (problem zidentyfikowany jako „problem wielu rąk").

## Znacząca kontrola człowieka. Uwarunkowania

Ryzyko szkód (i odpowiedzialność) związane z ewolucją robotów wynika z podstawowego problemu kontroli nad (semi) autonomicznymi systemami robotycznymi. Literatura z zakresu AI formułuje rozwiązania tego problemu w aspekcie **znaczącej kontroli człowieka** (*Meaningful Human Control*, MHC) (Santoni de Sio, van den Hoven, 2018). Zgodnie z tą zasadą to ludzie powinni ostatecznie kontrolować, a tym samym być moralnie odpowiedzialni za wszelkie decyzje podejmowane przez AI. W sytuacji, gdy człowiek nie ma bezpośredniej kontroli, może i powinien mieć możliwość kontroli pośredniej, pozwalającej na określenie odpowiedzialności (Nyholm, 2020).

W kontekście ewolucji robotów oznaczałoby to, że już na etapie projektowania podejmowane być muszą określone środki ostrożności. Najczęściej zalicza się do nich:

- scentralizowany system reprodukcji;
- zaawansowane systemy predykcyjne;
- programowanie wykluczające wybór celów niebezpiecznych dla ludzi (dodawanie wartości). Przykładowo: system można skonfigurować tak, aby roboty „nie chciały" się rozmnażać niezależnie (Eiben, Ellers, Meynen, Nyholm, 2021).

Istotne jest, by działania podejmowane w ramach tych środków ostrożności były na bieżąco dostosowywane do rozwoju sytuacji. W innym przypadku będą mało skuteczne. Ewoluujące roboty reprezentują bowiem zupełnie nowy rodzaj maszyn, które mogą zmieniać swoje zachowanie. Zdaniem autorów publikacji *Robot Evolution: Ethical Concerns* (Eiben, Ellers, Meynen, Nyholm, 2021) mogą (1) „uciec" przed wdrożonymi środkami kontroli, np. nie udostępniając swoich danych operacyjnych. Bardzo mało prawdopodobną, ale wyobrażalną drogą ucieczki jest scenariusz *Parku Jurajskiego*, w którym roboty znajdują alternatywny sposób reprodukcji poza centralnym ośrodkiem reprodukcji. Mogą również (2) chcieć wykorzystać głęboko zakorzenione w ludziach emocjonalne wzorce kontaktów z innymi ludźmi czy ze zwierzętami (wrażliwość, normatywne osądy itp.), osłabiając tym samym zdolność ludzkiego kontrolera do bycia obiektywnym w procesie ingerencji w ewolucję robota i ewentualnego przerwania procesu ich reprodukcji w sytuacji zagrożenia.

Zgodnie z koncepcją Filippo Santoniego de Sio i Jeroena van den Hovena (2018), aby systemy sztucznej inteligencji znajdowały się pod **znaczącą kontrolą człowieka** (MHC), muszą być spełnione dwa warunki zwane *tracking*[8]

[8] *Traking* wymaga, aby system społeczno-techniczny – tj. cała kombinacja elementów technicznych, ludzkich i organizacyjnych – był zaprojektowany w taki sposób, aby w widoczny sposób reagował na intencje odpowiednich agentów i na istotne fakty w środowisku.

i *tracing*[9]. Warunki te opisują charakter relacji oraz cechy, do jakich powinien dążyć system człowiek-maszyna, aby zachować odpowiedzialność człowieka. *Tracking* wymaga dostosowania systemu do wartości, powodów i intencji działań odpowiednich agentów ludzkich, a *tracing* wymaga dostosowania systemu do możliwości (umiejętności) ludzi. Zgodnie z operacyjno-przyczynowym pojęciem kontroli przyjętym w naukach technicznych system techniczny znajduje się pod kontrolą czynnika ludzkiego, gdy istnieje wiarygodny związek przyczynowy między nim a zachowaniem maszyny. Filozoficzna i normatywna idea stojąca za teorią znaczącej kontroli człowieka zakłada, że moralnie istotna kontrola i odpowiedzialność moralna zależą od dostosowania systemu socjotechnicznego do pobudek i możliwości odpowiednich czynników ludzkich (źródłem „sensowności" są pobudki i możliwości, a nie zachowanie).

W tabeli 2 przedstawiono schematycznie uwarunkowania i sposoby działania w przypadku, gdy celem jest wypełnienie opisanych w tabeli 1 luk odpowiedzialności.

Tabela 2. Znacząca kontrola ludzka w kontekście odpowiedzialności

| **Znacząca kontrola człowieka** | | |
|---|---|---|
| **Warunki** | **Działania** | **Wpływ na odpowiedzialność** |
| *Tracking*: dostosowanie systemu do wartości, powodów i intencji działań ludzkich agentów | Mapowanie działań agentów zaangażowanych w funkcjonowanie systemu (cele/zamierzenia) i ich relacji z systemem | **Odpowiedzialność bierna**<br>Rozszerza zakres agentów, jeżeli chodzi o przypisanie winy i odpowiedzialności (np. firmy i organy regulacyjne za wypadki z udziałem samochodów autonomicznych)<br><br>**Czynna odpowiedzialność**<br>Wspiera menedżerów, projektantów, decydentów, badaczy w identyfikowaniu potencjalnych napięć wartości i przyjmowaniu odpowiedzialności za tworzone projekty |

[9] W przeciwieństwie do propozycji opartych na nowych formach odpowiedzialności prawnej (Maternowska, 2022) koncepcja MHC proponuje, aby systemy społeczno-techniczne były systematycznie projektowane w celu uniknięcia luk w zakresie winy moralnej, odpowiedzialności biernej i czynnej. Warunek *tracing* sugeruje, że system może pozostać pod kontrolą MHC tylko w przypadku znacznego zrównoważenia systemu z technicznymi, motywacyjnymi i moralnymi zdolnościami odpowiednich agentów zaangażowanych w projektowanie, kontrolę i eksploatację systemu. Bezpośrednim celem tego warunku jest promowanie sprawiedliwego podziału winy moralnej, a tym samym uniknięcie dwóch niepożądanych skutków: po pierwsze, zjawiska „kozła ofiarnego", tj. pociągania do odpowiedzialności agentów bez odpowiedniej zdolności do uniknięcia wykroczenia; po drugie, bezkarności za spowodowanie wypadków, których można było uniknąć.

*cd. tab. 2*

| | | |
|---|---|---|
| *Tracing*: dostosowanie systemu do możliwości ludzkich agentów | Analiza możliwości ludzkich agentów w systemie oraz sprawdzenie, czy system wystarczająco je odzwierciedla | **Odpowiedzialność na zasadzie winy**<br>Pozwala na sprawiedliwe atrybucje; unikanie „kozła ofiarnego" i bezkarności w przypadku wypadków (obwinianie jedynie kierowcy za możliwy do uniknięcia wypadek samojezdnego samochodu)<br><br>**Odpowiedzialność moralna**<br>Nalega, aby odpowiedni agenci rozumieli procesy robocze AI i ich rolę w podejmowaniu decyzji (np. lekarz używający sztucznej inteligencji do diagnozy)<br><br>**Publiczna odpowiedzialność**<br>Wymaga, aby funkcjonariusze publiczni (nie tylko eksperci IT i firmy prywatne) byli zdolni i zmotywowani do działania pozwalającego zrozumieć procesy robocze AI i nakłada na nich obowiązek przekazywania informacji społeczeństwu<br><br>**Czynna odpowiedzialność**<br>Wspiera różne podmioty w rozwijaniu wiedzy, zdolności, możliwości i motywacji do wypełniania różnych obowiązków związanych ze sztuczną inteligencją |

Źródło: Santoni de Sio, Mecacci, 2021

W tym miejscu i trochę na marginesie dotychczasowych rozważań nasuwa się pytanie: co stanie się w sytuacji, gdy roboty, które mogą ewoluować i uczyć się, uzna się za formę sztucznego życia[10]?

Dotychczasowe rozważania dotyczyły w istocie ochrony rasy ludzkiej przed negatywnymi skutkami ewolucji robotów. Gdy jednak kwestię ochrony odwróci się (ochrona ewoluujących robotów przed ludźmi), od razu implikuje to innego rodzaju rozważania etyczne (Bryson, 2018). Kluczem jest postrzeganie populacji robotów jako gatunku, co wymaga uwzględnienia moralnego aspektu tego zagadnienia. Przyjęcie, że roboty są bytem – nie ważne, że sztucznym – pociąga za sobą konsekwencje postrzegane jako powstanie określonego obowiązku moralnego, podobnie jak funkcjonuje to w przypadku wielorybów, delfinów, wilków, psów i kotów itp. (Gordon, Nyholm, 2021). W myśl tego roboty, a wraz z nimi proces ich ewolucji, zasługiwałyby na określony poziom ochrony. To rodzi kwestię praw

[10] Zwolennicy tego poglądu tłumaczą go najczęściej dwoma argumentami: po pierwsze, w biologii kluczową różnicą między życiem a nie-życiem jest reprodukcja. Roboty mają możliwość reprodukcji i dzielą się z innymi formami życia cechami, takimi jak ruch i zużycie energii. Po drugie, roboty nie tylko potrafią się rozmnażać, lecz są na równi tworem ludzkiego projektu i procesu ewolucji.

robotów, analogicznie do istniejących praw zwierząt (Bertolini, 2020; Bertolini, Riccaboni, 2020). Biorąc to pod uwagę, można by zatem kwestionować niektóre interwencje, takie jak użycie „wyłącznika awaryjnego", i zastanawiać się, czy są one etyczne względem robotów jako form sztucznego życia. Istotne pytanie brzmi tak: czy zakończenie „życia" robotów ewolucyjnych powinno być postrzegane jako „wyłączenie maszyny" czy jako „zabijanie żywej istoty"?

W odniesieniu do analizowanego zagadnienia znaczącej kontroli człowieka można jednak stwierdzić, że w każdym przypadku tego typu względy moralne mogą ją potencjalnie ograniczyć.

## Podsumowanie

Zdaniem cytowanego wcześniej prof. Dragana (Wyżga, Siedzik, 2023) mamy obecnie do czynienia z największym cywilizacyjnym przewrotem, jaki kiedykolwiek nastąpił. Spowodował go gwałtowny rozwój sztucznej inteligencji i uczenia maszynowego.

Pojęcia ewolucji robotów i uczenia maszynowego są pokrewne. W przypadku maszyn postępy są tak szybkie, że zanika różnica między nauką nowych umiejętności a ewolucją przeobrażającą całą konstrukcję – zwłaszcza gdy robot uczy się samemu rozbudowywać i modyfikować własne „ciało" w celu lepszego wykonywania zadań[11]. Zatem ewolucja robotów nie jest już *science fiction.* Technologia rozwija się tak szybko, że prawdopodobnie pierwsze w pełni autonomicznie odtwarzające się i ewoluujące roboty pojawią się w ciągu dekady (Eiben, Ellers, Meynen, Nyholm, 2021). Obecnie badania w tej dziedzinie są zazwyczaj napędzane ciekawością, ale w przyszłości z pewnością będą coraz bardziej zorientowane na praktyczne zastosowania.

Rozwój tej technologii zwiastuje wiele nowych i trudnych problemów natury etycznej i prawnej i świadomość tego powinna pojawić się wcześniej, nim technologia stanie się dojrzała. Naukowcy i potencjalni użytkownicy już teraz powinni rozważać kwestie kontroli procesu ewolucji robotów. W szczególności dotyczy to opracowania wytycznych (etycznych i metodologicznych) do projektowania tego typu systemów w celu minimalizacji potencjalnych szkód i maksymalizacji korzyści.

Choć proces ewolucyjny jest autonomiczny, to nadal wiąże się z koniecznością przypisania odpowiedzialności. Ma to kluczowe znaczenie nie tylko w odniesieniu do egzekwowania odpowiedzialności w przypadku niepowodzenia (powstania

[11] Naukowcy z ETH w Zurichu i z Uniwersytetu w Cambridge stworzyli system robotyczny, który jest zdolny do „ewolucji" i doskonalenia samego siebie. W praktyce wygląda to w ten sposób, że duże ramię robota buduje swoje „dzieci", czyli mniejsze urządzenia, które następnie obserwuje i na podstawie tych obserwacji wnioskuje, co trzeba poprawić w kolejnej generacji robotów (Usidus, 2023).

szkody), lecz także zapewnienia, że ludzie wezmą odpowiedzialność za pewne aspekty tego procesu. Bez przyjmowania odpowiedzialności przez ludzi proces ewolucji robotów nie może być skutecznie kontrolowany (Eiben, Ellers, Meynen, Nyholm, 2021). Przyszłe prace filozoficzne, empiryczne i techniczne będą musiały między innymi dokładniej wyjaśnić, jakie luki w zakresie odpowiedzialności mogą pojawić się w odniesieniu do różnorodnych systemów, w różnych kontekstach zastosowania oraz w jakim stopniu podejście do przedstawionej w artykule koncepcji znaczącej kontroli człowieka może skutecznie wypełnić te luki.

BIBLIOGRAFIA

[1] Amoroso, D., Tamburrini, G., 2019. *What makes human control over weapon systems „meaningful"?*, https://www.icrac.net/wp-content/uploads/2019/08/Amoroso-Tamburrini_Human--Control_ICRAC-WP4.pdf (dostęp: 27.03.2023).

[2] Ashlock, D., 2006. *Evolutionary Computation for Modeling and Optimization*, New York: Springer, https://theswissbay.ch/pdf/Gentoomen%20Library/Artificial%20Intelligence/Evolutionary%20computation/Evolutionary%20Computation%20for%20Modeling%20and%20Optimization%20-%20Daniel%20Ashlock.pdf (dostęp: 27.03.2023).

[3] Badyaev, A.V., 2011. *Origin of the Fittest: Link between Emergent. Variation and Evolutionary Change as a Critical Question in Evolutionary Biology*, https://royalsocietypublishing.org/doi/10.1098/rspb.2011.0548 (dostęp: 24.02.2023).

[4] Bertolini, A., 2020. *Artificial Intelligence and Civil Liability*, Bruxelles, https://www.europarl.europa.eu/RegData/etudes/STUD/2020/621926/IPOL_STU(2020)621926_EN.pdf (dostęp: 27.03.2023).

[5] Bertolini, A., Riccaboni, M., 2020. Grounding the Case for a European Approach to the Regulation of Automated Driving: The Technology-Selection Effect of Liability Rules, *European Journal of Law and Economics*, nr 51(27).

[6] Bongard, J.C., 2013. Evolutionary Robotics, *Communications of the ACM*, nr 56, https://pages.vassar.edu/evodevorobotics/files/2014/10/Bongard-2013.pdf (dostęp: 27.03.2023).

[7] Bryson, J.J., 2018. Patiency Is Not a Virtue: the Design of Intelligent Systems and Systems of Ethics, *Ethics and Information Technology*, nr 20(1).

[8] Čapek, K., 1920. *R.U.R.: Rossum's Universal Robots* (English Translation), https://www.gutenberg.org/files/59112/59112-h/59112-h.htm (dostęp: 24.02.2023).

[9] Danaher, J., 2020. Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism, *Science and Engineering Ethics*, nr 26(4).

[10] de Jong, K.A., 2006. *Evolutionary Computation: A Unified Approach*, Cambridge: MIT Press, https://www.researchgate.net/publication/220691482_Evolutionary_Computation_A_Unified_Approach (dostęp: 24.02.2023).

[11] Doncieux, S., Bredeche, N., Mouret, J.-B., Eiben, A.E., 2015. *Evolutionary Robotics: what, why, and where to*, https://www.frontiersin.org/articles/10.3389/frobt.2015.00004/full (dostęp: 24.02.2023).

[12] Eiben, A.E., Ellers, J., Meynen, G., Nyholm, S., 2021. *Robot Evolution: Ethical Concerns*, https://www.frontiersin.org/articles/10.3389/frobt.2021.744590/full (dostęp: 27.03.2023).

[13] Eiben, A.E., Smith, J.E., 2003. *Introduction to Evolutionary Computing*, https://warin.ca/ressources/books/2015_Book_IntroductionToEvolutionary Comp.pdf (dostęp: 27.03.2023).

[14] Ellery, A., 2020. How to Build a Biological Machine Using Engineering Materials and Methods, *Biomimetics (Basel)*, nr 5(3).

[15] Ellery, A., Eiben, A.E., 2019. *To Evolve or Not to Evolve? That Is the Question*, https://digitalcollection.zhaw.ch/bitstream/11475/17907/1/ALIFE_2019.pdf (dostęp: 27.03.2023).

[16] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, Ch., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., Vayena, E., 2018. *AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*, https://link.springer.com/article/10.1007/s11023-018-9482-5 (dostęp: 27.03.2023).

[17] Gordon, J.S., Nyholm, S., 2021. *Ethics of Artificial Intelligence*, Internet Encyclopedia of Philosophy, https://iep.utm.edu/ethics-of-artificial-intelligence/ (dostęp: 27.03.2023).

[18] Kriegman, S., Blackiston, D., Levin, M., Bongard, J., 2020. A Scalable Pipeline for Designing Reconfigurable Organisms, *Proceedings of the National Academy of Sciences of the United States of America*, nr 117.

[19] Lipson, H., Pollack, J.B., 2000. *Automatic Design and Manufacture of Robotic Lifeforms*, https://www.cs.swarthmore.edu/~meeden/DevelopmentalRobotics/lipson00.pdf (dostęp: 24.02.2023).

[20] Maternowska, M., 2019. *Nowe technologie i ich wpływ na łańcuchy dostaw: sztuczna inteligencja*, https://www.ue.katowice.pl/fileadmin/user_upload/wydawnictwo/SE_Artyku%C5%82y_381_400/SE_388/04.pdf (dostęp: 27.03.2023).

[21] Maternowska, M., 2022. *Dylematy odpowiedzialności za roboty sterowane sztuczną inteligencją*, https://nsz.wat.edu.pl/pdf-155318-82214?filename=Dylematy.pdf (dostęp: 27.03.2023).

[22] Matthias, A., 2004. The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata, *Ethics and Information Technology*, nr 6(3).

[23] Nolfi, S., Floreano, D., 2000. *Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organizing Machines*, https://www.researchgate.net/publication/226668077_Evolutionary_Robotics (dostęp: 27.03.2023).

[24] Nyholm, S., 2020. *Humans and Robots: Ethics, Agency, and Anthropomorphism*, London: Rowman & Littlefield International.

[25] Santoni de Sio, F., Mecacci, G., 2021. *Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them*, https://link.springer.com/article/10.1007/s13347-021-00450-x (dostęp: 27.03.2023).

[26] Santoni de Sio, F., van den Hoven, J., 2018. Meaningful Human Control over Autonomous Systems: A Philosophical Account, *Frontiers in Robotics and AI*, nr 5.

[27] Usidus, M., 2023. *Ewolucja robotów*, https://mlodytechnik.pl/technika/29780-ewolucja-robotow (dostęp: 27.03.2023).

[28] van de Poel, I., 2016. An Ethical Framework for Evaluating Experimental Technology, *Science and Engineering Ethics*, nr 22(3).

[29] Vargas, P.A., di Paolo, E.A., Harvey, I.M., Husbands, P., Moioli, R., 2014. *The Horizons of Evolutionary Robotics*, https://terrorgum.com/tfox/books/horizonsofevolutionaryrobotics.pdf. (dostęp: 27.03.2023).

[30] Wyżga, P., Siedzik, D., 2023. *Zapowiada koniec świata, jaki znamy. Polski fizyk o rewolucji AI*, https://wiadomosci.wp.pl/polski-fizyk-pochwalil-sie-odkryciem-teraz-mowi-o-swiatowej-rewolucji-6881287875992384a (dostęp: 2.04.2023).