AVANT, wol. XVI, nr 3, ISSN 2082-6710
https://doi.org/10.26913/ava32502

Podmiotowość refleksyjna a autonomia sensomotoryczna w kontekście sprawczości sztucznych systemów poznawczych

Barbara Tomczyk
Katedra Logiki i Kognitywistyki
Uniwersytet Marii Curie-Skłodowskiej w Lublinie
barbara.tomczyk@mail.umcs.pl

Przyjęto 2024-11-12; zaakceptowano 2025-07-13; opublikowano *Online First* 2025-08-02.

Abstrakt

W prezentowanym artykule przywołuję dwa pojęcia sprawczości: pierwsze sformułowano w ramach analitycznej filozofii działania intencjonalnego, zakładającej refleksyjność podmiotu, a drugie rozwinięto na gruncie enaktywistycznej koncepcji sprawczości jako autonomii sensomotorycznej. Te dwie koncepcje autonomicznego działania umieszczam w kontekście dyskusji na temat sprawczości sztucznych systemów, które obecnie zyskują coraz większą niezależność, jeśli chodzi o dobór metod rozwiązywania konkretnych zadań. Polemizując ze swoimi wcześniejszymi argumentami wyrażonymi w książce *Podmiotowość rozszerzonych systemów poznawczych* (2022), wskazuję na przewagę eksplanacyjną koncepcji enaktywistycznej, która, mimo że nie jest w stanie spełnić niektórych funkcji koncepcji refleksyjnej, jest bardziej użyteczna w dyskusjach nad moralnymi konsekwencjami rozwoju sztucznej inteligencji. Konsekwencje te dyskutowane są obecnie żywo w ramach problemu uzgodnienia wartości ludzkich z tymi, którymi kieruje się, a w szczególności będzie się kierowała w przyszłości, sztuczna inteligencja (ang. *The Value Alignment Problem*). W pracy wskazuję również na zastosowania enaktywistycznej koncepcji sprawczości w wyjaśnianiu funkcjonowania systemów grupowych jako sprawców działań.

Słowa kluczowe: autonomia, sprawczość, podmiot działania, enaktywizm, sztuczna inteligencja, autopojeza

1. Wstęp

Rozwój technologii sztucznej inteligencji, który obserwujemy obecnie, prowadzi do powstania nowego typu systemów, które rozwiązują zadania poznawcze i jako takie są sprawcami działań. Jednak zarówno pojęcie działania, jak i jego podmiotu, są przedmiotami rozważań prowadzonych od wieków przez filozofów, a ich efekty często wskazują na człowieka jako na jedyny obiekt mogący przejawiać podmiotowość sprawczą. Czy zatem stosowanie owych pojęć do sztucznych systemów nie jest nadużyciem lub określeniem metaforycznym? Czy należy ukuć nowe pojęcia sprawczości i działania, które będą określały to, co robi sztuczna inteligencja? Czy tylko człowiek może być autentycznym sprawcą? Odpowiedzi na te pytania zależą od przyjętej koncepcji autonomicznego działania. W prezentowanej pracy wskazuję na dwie z nich: refleksyjną – sformułowaną w ramach analitycznej filozofii działania intencjonalnego, oraz sensomotoryczną – rozwiniętą na gruncie enaktywistycznej koncepcji sprawczości. Koncepcje te mają odmienne zastosowania i okazują się owocne w innych obszarach refleksji. Pierwsza z nich ma tę przewagę, że umożliwia analizę pojęcia odpowiedzialności podmiotu za podjęte działanie, druga natomiast ujawnia swą moc eksplanacyjną w rozważaniach nad aktywnością systemów pozaludzkich.

Tradycja filozoficznego namysłu nad podmiotowością przeniknięta jest ideą wyjątkowości ludzkiego poznania i działania ugruntowaną w refleksyjnej koncepcji podmiotu. Użycie do jej określenia pojęcia „refleksyjności" uzasadniam tym, że filozofowie działania pracujący w jej ramach odwołują się do takich cech podmiotu jak dostęp do treści stanów umysłowych, doświadczenie bycia ich podmiotem, odczucie posiadania wolnego wyboru pomiędzy alternatywnymi działaniami i związana z emocjami wrażliwość na wartości. Podmiot refleksyjny cechuje się również jaźnią, rozumianą jako posiadanie subiektywnego punktu widzenia gwarantującego jego psychiczną tożsamość w czasie (Bremer, 2016). To dzięki dysponowaniu pojęciem „ja" człowiek staje się świadomy swojej podmiotowości i wypracowuje inne pojęcia, na których wspiera się życie społeczne i kultura, takie jak „dobro", „zło", „godność", „wolność", czy „miłość" (Taylor, 2001).

Mimo że wspomniane powyżej cechy podmiotu refleksyjnego może przejawiać obecnie tylko człowiek, nie ma żadnych pojęciowych przeciwskazań, aby inne systemy również osiągnęły poziom złożoności, który to umożliwia. Czy można jednak uzyskać taką złożoność w sposób sztuczny? Jaką metodę obrać, aby podjąć taką próbę, i po co właściwie to robić? W pracy tej argumentuję, że owocną ramę pojęciową do odpowiedzi na te pytania tworzy druga przywołana przeze mnie koncepcja podmiotowości, koncepcja, która powstała na gruncie enaktywizmu i która definiuje sprawczość w terminach autonomii sensomotorycznej. Moja argumentacja będzie zmierzać do wykazania, że posiada ona przewagę eksplanacyjną nad ujęciami refleksyjnymi, która ujawnia się na gruncie rozwa-

żań nad efektami rozwoju sztucznej inteligencji, a także nad sprawczością systemów grupowych. W tym zakresie prezentowany artykuł podejmuje polemikę z argumentacją znajdującą się w mojej książce *Podmiotowość rozszerzonych systemów poznawczych* (Tomczyk, 2022), w której wyrażam przekonanie, że autentyczną podmiotowość można przypisać tylko i wyłącznie systemowi refleksyjnemu, a zatem musi ona zostać zredukowana w systemach rozszerzonych do człowieka. W obecnej pracy argumentuję, że enaktywizm prezentuje szerszą perspektywę refleksji nad sprawczością, dzięki której można dostrzec, że podmiotowość refleksyjna wyłania się z bardziej podstawowych form sprawczości. Skłania to do wniosku, że jeśli sztuczne systemy spełnią enaktywistyczne warunki sprawczości, mogą również osiągnąć – na pewnym etapie rozwoju swojej złożoności – refleksyjność podobną do ludzkiej. Celem tego artykułu nie jest jednakże argumentacja na rzecz możliwości przypisywania refleksyjności systemom pozaludzkim, lecz wskazanie na alternatywne pojęcie autonomicznego sprawstwa i jego moc eksplanacyjną w badaniach nad wyłanianiem się coraz bardziej złożonych form sprawczości. W takiej perspektywie podmiotowość refleksyjna nie prezentuje się jako ekskluzywna, nowa jakość stanowiąca istotę ludzkiej natury, lecz jako efekt wysoce złożonej sensomotorycznej aktywności podmiotu ucieleśnionego i zakorzenionego w świecie.

Artykuł rozpoczynam od wskazania głównych warunków podmiotowości refleksyjnej. W pierwszej części przytaczam również argumenty przeciwko ograniczaniu sprawczości do elementu ludzkiego w rozszerzonych systemach poznawczych. Następnie, w drugiej części, prezentuję enaktywistyczną koncepcję autonomii sensomotorycznej po to, aby w trzeciej części wskazać na próby zastosowania jej do projektowania sztucznych systemów poznawczych. W części tej podejmuję również refleksję nad istotnymi różnicami pomiędzy żywymi a sztucznymi systemami poznawczymi, przytaczam obecne w literaturze warunki autonomicznego, w sensie enaktywistycznym, sztucznego systemu oraz wskazuję na najpoważniejsze bariery stojące na drodze do jego stworzenia. Pojawia się w tym miejscu pytanie, w jakim celu właściwie podejmować wysiłki zmierzające do upodobnienia sztucznej inteligencji do ludzkiej. Odpowiadam na nie w ostatniej części pracy, wskazując na problem uzgodnienia wartości (ang. *The Value Alignment Problem*), który jest żywo dyskutowany na gruncie etyki maszynowej i filozofii moralności. Artykuł kończę refleksją nad przewagą eksplanacyjną enaktywistycznej koncepcji sprawczości w rozważaniach nad podmiotowością systemów grupowych.

2. Autonomiczny sprawca jako istota refleksyjna

W książce *Podmiotowość rozszerzony systemów poznawczych* rozważałam, jakie warunki powinien spełnić system poznawczy, aby stanowił podmiot poznania, wiedzy i działania, a co za tym idzie, podmiot odpowiedzialności epistemi-

cznej i moralnej. Przedmiotem moich analiz były dwa rodzaje systemów: rozszerzony system człowiek–artefakt oraz system grupowy. Po analizie różnych koncepcji podmiotowości i ich implikacji, zarówno teoretycznych, jak i praktycznych, sformułowałam wniosek, że człowiek stanowi w takich systemach element wyróżniony w kontekście sprawczości. Ten wyjątkowy status wiązałam z takimi jego cechami jak refleksyjność, decyzyjność, możliwość podjęcia świadomego wyboru pomiędzy różnymi działaniami w oparciu o posiadaną hierarchię wartości oraz wzięcia za ten wybór odpowiedzialności związanej z pochwałą bądź potępieniem moralnym. W prezentowanym artykule mam zamiar spojrzeć na problem sprawczości pod innym kątem, z punktu widzenia stanowisk poszerzających granice podmiotowości poznawczej i sprawczej w sposób, który umożliwia ujęcie w zbiorze podmiotów nie tylko innych żywych istot poza człowiekiem, lecz również sztucznych systemów poznawczych. Zanim jednak przejdę do argumentacji na rzecz efektywności stanowiska sensomotorycznego w rozwiązywaniu filozoficznych wyzwań związanych z rozwojem współczesnej technologii, wskażę na główne racje za przypisywaniem autentycznej podmiotowości poznawczej i sprawczej wyłącznie istocie refleksyjnej, jaką jest człowiek. Pozwoli to na jasne określenie różnic pomiędzy dwoma pojęciami sprawczości: refleksyjnej i sensomotorycznej.

Filozoficzna debata na temat warunków autonomicznego działania jest bardzo intensywna i uczestniczą w niej przedstawiciele wielu stanowisk[1]. W książce *Podmiotowość rozszerzonych systemów poznawczych* odwoływałam się głównie do koncepcji, które łączy teza, że u podstaw autonomicznego działania leżą: zdolność do rozumowania, posiadanie postaw propozycjonalnych, takich jak przekonania, pragnienia i intencje, które determinują konkretne działanie, oraz znajomość i możliwość oceny ich treści. Wśród przedstawicieli tych stanowisk można wymienić zarówno internalistów, jak i eksternalistów[2]. Takie rozumienie sprawczości obecne jest również w analitycznej filozofii działania, reprezentowanej między innymi przez Donalda Davidsona. Argumentuje on, że o działaniu może być mowa tylko wtedy, gdy jego podmiot ma zasadniczo dostęp do treści stanów umysłowych stanowiących zarówno przyczyny podejmowanych działań, jak i racje w procesie ich uzasadniania (1967/2002). Sprawca w tradycji filozoficznej jawi się zatem jako system refleksyjny,

[1] Przedstawienie najważniejszych stanowisk wobec autonomicznego działania znajduje się w moim artykule „The Locus of Agency in Extended Cognitive Systems" (Tomczyk, 2023).

[2] Przedstawiciele stanowisk internalistycznych wobec autonomicznego działania skupiają się na relacjach między postawami propozycjonalnymi podmiotu, jego zdolnościami do wnioskowania na ich podstawie i do racjonalnej refleksji nad ich spójnością (Dworkin, 1988; Frankfurt, 1971; Bratman, 1979). Eksternaliści przekonują natomiast, że dla oceny autonomiczności działania bardziej istotne jest to, aby podmiot był świadomość tego, dlaczego dane postawy są odpowiednimi racjami za określonym działaniem (Fischer i Ravizza, 1993), lub czy proces prowadzący do sformułowania owych racji był rzetelny (nie był manipulacją, indoktrynacją itd.) (Christman, 1991). Więcej na ten temat w: Tomczyk, 2023.

inicjujący działanie na zasadzie wolnego wyboru spośród alternatyw, system, który jest intencjonalny w tym sensie, że procesy poznawcze odsłaniają mu świat, czyli skutkują reprezentacjami posiadającymi treść dostępną dla niego i zrozumiałą.

Niewątpliwie, aby działać autonomicznie, system musi posiadać moc sprawczą. Pytanie, czym ona jest stanowi jeden z podstawowych problemów rozważanych przez badaczy zjawiska autonomicznego działania (Chisholm, 1976; Clarke i Reed ,2015; O'Connor, 2009; Cash, 2010; Christman, 2004). Ogólnie panuje zgodna co do tego, że źródłem mocy sprawczej inicjującej działanie musi być sam podmiot, a gdy pochodzi ona spoza niego, nie stanowi on sprawcy w danym przypadku. W jakim stopniu jednak podmiot powinien posiadać kontrolę nad czynnikami, które wpływają na jego wybory, aby działanie można było nazwać autonomicznym? Jest to kwestia sporna, a jej rozstrzygnięcie na gruncie koncepcji odwołujących się do refleksyjności, kontroli, świadomości i wolności wydaje się niemożliwe ze względu na wieloznaczność tych pojęć i mnogość wyjaśniających je koncepcji. Z tego między innymi powodu alternatywne ujęcie autonomii jako sprawczości sensomotorycznej kusi jasnością i stabilnością, chociaż odbiega znacznie od pierwotnych intencji przyświecających filozofom analizującym pojęcie działania.

Badacze wiążący autentyczną sprawczość z refleksyjnością zwracają również uwagę na warunek, którym jest poczucie sprawczości doświadczane przez podmiot działania. Odczuwanie kontroli, inicjatywy, wolnego wyboru i wysiłku z nim związanego uznawane są za konieczne cechy sprawcy. Bez nich podmiot nie byłby w stanie sformułować przekonań i pragnień uzasadniających działanie i umożliwiających przypisanie mu odpowiedzialności za nie (Bayne, 2008; Nida-Rümelin, 2018). Określenie, czym jest owo doświadczenie sprawczości i jak ocenić jego zachodzenie z perspektywy trzecioosobowej w celu przypisania go systemom, które nie są w stanie go raportować, jest podobnie problematyczne jak zagadnienie źródła mocy sprawczej i stopnia kontroli nad czynnikami determinującymi wybory działania.

Odwołując się do wskazanych powyżej warunków refleksyjnej sprawczości, można bronić tezy, że podmiotowość rozszerzonego systemu poznawczego złożonego z człowieka i artefaktu jest zredukowana do elementu ludzkiego (Tomczyk, 2022). System taki, krótko przypominając, konstytuowany jest przez ciągłą wzajemną relację przyczynową (ang. *continuous reciprocal causation*) między jego elementami, w której nie sposób rozróżnić ludzkich procesów poznawczych od procesów zachodzących w artefakcie ze względu na ich ścisłe sprzężenie i wzajemną zależność (Clark, 2008). Przykładami artefaktów, które mogą tworzyć z ludźmi takie sprzężone systemy, są smartfony i komputery z ich różnymi aplikacjami, urządzenia substytucji sensorycznej lub interfejsy mózg-komputer (BCI). Podczas gdy procesy poznawcze są rozproszone w takim przypadku na cały rozszerzony system, podmiotem tych procesów, jak również stanów umysłowych będących ich efektem oraz działań wykonywanych na ich

podstawie, jest człowiek, a nie człowiek+artefakt. Wyłącznie człowiek spełnia bowiem warunki autentycznej sprawczości, wyróżnione przez filozofów wskazujących na refleksyjność jako jej cechę konieczną.

Niektórzy filozofowie argumentują jednak, że artefakt w rozszerzonym systemie poznawczym nie powinien być traktowany jako element zewnętrzny wobec ludzkiej podmiotowości, lecz jako czynnik współtworzący nową jednostkę poznającą i działającą, którą pod względem teoretycznym i praktycznym powinno się analizować jako odrębnego sprawcę (Farina i Lavazza, 2022; 2022a). Zdaniem Mirko Fariny i Andrea Lavazzy, taką strategię wyjaśniającą powinno się promować ze względów moralnych, gdyż w jej ramach nie traktuje się artefaktów jako zagrożenia dla autonomii i samodeterminacji jednostki, którego widmo stanowi barierę dla ich społecznej akceptacji. Artefakty mianowicie nie stanowią zagrożenia dla ludzkiej sprawczości opartej na wolnym wyborze i samokontroli, gdyż współtworzą sprawcę, którym jest sztucznie wzmocniony człowiek, na zasadzie funkcjonalnej równości z naturalnymi procesami biologicznymi zachodzącymi w jego organizmie. Podmiot nie powinien być zatem ograniczony do biologicznego organizmu, lecz rozszerzony na cały system człowiek+artefakt. U podstaw tezy, że taka interpretacja posiada przewagę moralną nad ujęciem redukującym sprawczość do niewzmocnionego człowieka, leży założenie o pozytywnym wpływie sztucznych wzmocnień poznawczych na rozwój jednostek i w efekcie na rozwój społeczny. Rozszerzony system poznawczy, lub mówiąc inaczej, sztucznie wzmocniony człowiek (cyborg), jest jednym słowem bardziej efektywny poznawczo od biologicznego organizmu, podejmuje bardziej racjonalne decyzje i jest skuteczniejszy w ich realizacji. Rzecz jasna, kompetencje poznawcze nie gwarantują wzrostu motywacji do działania moralnego, same owe motywacje mogą być jednak, po spełnieniu odpowiednich warunków, sztucznie wzmacniane przez współczesną technologię (Giubilini i Savulescu, 2017; Lara i Deckers, 2020). Zagadnienie sztucznych wzmocnień moralnych stanowi temat ożywionej dyskusji na gruncie współczesnej etyki i filozofii sztucznej inteligencji, w tym artykule jednak nie będę się w nie zagłębiać. Na potrzebę obecnej argumentacji wystarczy wskazać, że po spełnieniu odpowiednich warunków, przede wszystkim warunku świadomego włączenia artefaktu do charakteru poznawczego człowieka, artefakt staje się jego częścią, a co za tym idzie, nie stanowi zagrożenia dla jego samokontroli, autonomicznej sprawczości i odpowiedzialności. Co więcej, artefakt taki powinien być prawnie chroniony przed uszkodzeniem, w podobny sposób jak biologiczne części ludzkiego organizmu (Carter i Palermos, 2016).

Argumentację Fariny i Lavazzy uważam za przekonującą. Można z niej wyprowadzić konsekwencję głoszącą, że człowiek nie stanowi wyróżnionej pod względem podmiotowości poznawczej, sprawczej czy moralnej części rozszerzonego systemu poznawczego. Wszystkie zakładane przez badaczy autentycznej sprawczości cechy, takie jak intencjonalność, refleksyjność, moc sprawcza, czy doświadczenie sprawczości są konstytuowane przez działanie artefaktu.

Nie sposób zatem wyróżnić w tym systemie niewzmocnionego człowieka jako elementu przejawiającego autentyczną sprawczość, a ponadto nie powinniśmy tego robić, gdyż zubożamy w ten sposób człowieka o pewne jego zdolności poznawcze, które są dla niego transparentne i oczywiste w podobny sposób jak procesy biologiczne. W przypadku mianowicie, gdy sztucznie wzmocnione kompetencje poznawcze zawiodą, człowiek powinien zostać potraktowany pod względem odpowiedzialności za popełniony czyn tak, jakby zawiodły jego naturalne kompetencje, a nie jakby posiadał w pełni sprawny organizm[3].

Powyższe rozważania dotyczące teoretycznych i praktycznych implikacji stosowania sztucznych wzmocnień poznawczych prowadzą do wniosku, że autonomiczny sprawca nie zawsze sprowadza się do biologicznego organizmu, lecz może stanowić system złożony z takiego organizmu i artefaktu. Nie znaczy to, rzecz jasna, że odpowiedzialność za działanie spoczywa na całym systemie, co mogłoby skutkować takimi niedorzecznymi sytuacjami, w których Otto+notatnik idzie do więzienia[4]. Wciąż to konkretny człowiek staje przed sądem moralnym lub epistemicznym, rolą sędziów jednak jest wzięcie pod uwagę tego, co składa się na jego podmiotowość sprawczą, mianowicie czy jako sprawca stanowi on tylko biologiczny organizm, czy też sprzężony z artefaktem system. Niezależnie jednak od tego, czy rozważana jest sprawczość rozszerzonego systemu poznawczego, czy człowieka jako biologicznego organizmu, traktowana jest ona wciąż jako silnie związana z refleksyjnością. Założenie ciągle pozostaje to samo: tylko system, który jest świadomy swojej hierarchii wartości, treści przekonań, pragnień i intencji prowadzących do działania, może podejmować wolne decyzje, za które bierze odpowiedzialność. Owo powiązanie sprawczości z odpowiedzialnością i jej świadomym doświadczeniem jest ujęciem, które można nazwać maksymalistycznym, gdyż w jego ramach sprawcą może być tylko człowiek i do tego nie każdy.

W kolejnej części artykułu przyjrzę się alternatywnej koncepcji autonomicznego działania, którą można nazwać minimalistyczną, gdyż na jej gruncie przypisuje się sprawczość każdemu systemowi, który wykazuje cechy samowytwarzania (autopojezy) i adaptacyjności. Następnie wskażę, w jaki sposób sta-

[3] Farina i Lavazza ilustrują tę sytuację przykładem osoby, która używa naszyjnika z sensorem chemicznym, wydającego nieprzyjemny dźwięk, gdy poziom alkoholu w jej krwi przekroczy określony poziom, co chroni ją przez zachowaniem, które ona sama uznaje za moralnie naganne (Farina i Lavazza, 2022a). Przypuśćmy, że naszyjnik ten przestaje nagle działać, a osoba go nosząca upija się i powoduje wypadek. W tej sytuacji z osoby tej powinna zostać zdjęta część odpowiedzialności za skutki tego wypadku, tak jak jest ona zdejmowana z oso-by, której dysfunkcji uległy biologiczne procesy poznawcze, na przykład w skutek nieświadomego zażycia substancji powodującej spadek poczucia kontroli nad ilością spożytego alkoholu. Jednym słowem, bez owego naszyjnika osoba ta stanowi inny podmiot moralny niż w połączeniu z nim.

[4] Na absurdalne konsekwencje rozszerzenia umysłu Ottona na jego notatnik zwraca uwagę między innymi Erick Olson (2011).

nowisko to może zostać wykorzystane w aktualnej dyskusji dotyczącej sprawczości sztucznych systemów poznawczych i zagrożeń związanych z ich rosnącą autonomią. Z perspektywy enaktywistycznej koncepcji autonomii sensomotorycznej podmioty refleksyjne stanowią tylko niewielki podzbiór ogromnego zbioru autonomicznych sprawców.

3. Sprawczość jako autonomia sensomotoryczna

Enaktywistyczna koncepcja sprawczości oferuje wyjście poza antropocentryzm stanowisk refleksyjnych i ujęcie w zbiorze podmiotów działania systemów, które nie dysponują językiem pojęciowym ani poczuciem kontroli nad wyborem działania. Strategia taka skutkuje oderwaniem sprawczości od odpowiedzialności, co stanowi tak radykalny krok, iż pojawia się wątpliwość, czy wciąż uzasadnione jest stosowanie tego samego rozumienia „sprawstwa". Z drugiej strony, enaktywizm odwołuje się do podstawowej i koniecznej cechy systemu sprawczego, mianowicie do autonomicznego dążenia do utrzymania się w istnieniu. Motywacją tego stanowiska jest zatem wyróżnienie spośród wszystkich bytów tych, które przejawiają to dążenie, gdyż to właśnie ono decyduje o autonomii sprawczej. To, którym z owych systemów można przypisać winę lub zasługę jest dalszą kwestią stanowiącą przedmiot odrębnych analiz. Podmiot odpowiedzialności jest jednak wysoce złożonym podmiotem procesów adaptacyjnych, próbując zatem wykorzystać pojęcie sprawczości do systemów pozaludzkich, należy zacząć od analizy podstawowych warunków wszelkiego autonomicznego działania.

Z koncepcją sprawczości sensomotorycznej wiązane są nadzieje, które dużo trudniej zrealizować na gruncie stanowisk refleksyjnych. Dotyczą one możliwości przypisania sprawczości sztucznym systemom wykonującym zadania poznawcze. Motywacją ku temu jest coraz większa autonomia takich systemów polegająca na „podejmowaniu decyzji" bez bezpośredniej kontroli człowieka, a także chęć ludzi, aby decyzje te były zgodne z ich intencjami i celami. Aby to dążenie miało szansę na realizację, wartości, którymi kieruje się sztuczny podmiot działania, powinny przypominać te, ku realizacji których dążą ludzie. Wśród badaczy tej problematyki problem ten zyskał nazwę *The Value Alignment Problem* (dalej: problem uzgodnienia wartości) (Christian, 2020; Gabriel, 2020; Cannon, 2022) i zawiera się w pytaniu, w jaki sposób uzgodnić wartości sztucznej inteligencji z wartościami ludzkimi – po to, aby zapobiec ewentualnemu zagrożeniu z jej strony dla ludzkiego istnienia. Założenie przy tym jest takie, że niezależnie od tego, jakie wartości zostałyby zaimplementowane systemowi, bez zrozumienia ich „ducha", a nie tylko „litery" system może podjąć decyzje zgubne dla gatunku ludzkiego, eliminując go na przykład w celu rozwiązania problemu zmiany klimatycznej. Aby temu przeciwdziałać, należy sprawić, aby to, co jest ważne dla człowieka, było również ważne dla sztucznej inteligencji, innymi słowy, aby była ona podobna pod względem

sprawczości do człowieka. Stanowiska refleksyjne, odwołujące się do świadomego doświadczenia kontroli, sprawczości i podejmowania wolnej decyzji, nie stanowią w tym przedsięwzięciu dobrego punktu startu, gdyż już na wstępie stawiają bardzo wysokie wymagania; aby je spełnić, system musi najpierw osiągnąć autonomię rozumianą minimalistycznie, tak jak ją określają przedstawiciele enaktywizmu.

Enaktywizm wziął swoją nazwę od angielskiego słowa *enaction*, które odnosi się do dynamicznej sensomotorycznej aktywności konstytuującej procesy poznawcze podmiotu i stanowi opozycję wobec określenia *computation*, które jest stosowane do wyjaśniania tych procesów na gruncie obliczeniowych stanowisk kognitywistycznych. Podmiot poznający jest mianowicie rozumiany jako istota współtworząca poznawane przez siebie środowisko poprzez podejmowaną w nim aktywność (Varela, Thompson i Rosch, 1991; Noë, 2004). Zarówno percepcja, jak i bardziej złożone doświadczenia, takie jak poczucie własnej sprawczości, wyłaniają się z interakcji pomiędzy aktywnym organizmem a elementami jego środowiska, a interakcja ta jest tak ścisła, że nie sposób wyróżnić jej wewnętrznych (zachodzących w organizmie) i zewnę-trznych komponentów.

Według przedstawicieli enaktywizmu, aby być autonomicznym sprawcą, system musi być przede wszystkim autopojetyczny, co oznacza, że sam wyróżnia się z otoczenia, sam utrzymuje się w istnieniu i sam się regeneruje. Aby być do tego zdolnym, system musi przejawiać organizacyjne domknięcie (ang. *organizational closure*), czyli stanowić dynamiczną jedność złożoną ze sprzężonych ze sobą przyczynowo procesów, które wyznaczają jego zachowanie prowadzące do zdobywania w środowisku środków koniecznych do przetrwania (Varela, 1979; Maturana i Varela, 1980; 1992). Elementy środowiska inicjują zmiany w strukturze takiego systemu, lecz nie kontrolują bezpośrednio jego zachowania, które wyłania się z działania owej wewnętrznej organizacji. Owo pierwotne, autopojetyczne rozumienie biologicznej autonomii, wprowadzone przez Francisco Varelę i Humberto Marturanę, zostało uzupełnione przez późniejszych enaktywistów o warunek adaptacyjności i rozwinięte do postaci autonomii sensomotorycznej (Di Paolo, 2009; Beer i Di Paolo, 2023). Zdaniem Ezequiela Di Paolo, sama autopojetyczność nie jest wystarczającym warunkiem sprawczości, może bowiem istnieć system, który zachowuje swoją tożsamość i sam utrzymuje się w istnieniu w niezmiennym otoczeniu, lecz nie posiada żadnego mechanizmu umożliwiającego mu przetrwanie, gdy jego sytuacja w tym otoczeniu ulegnie zmianie i należy ją poprawić. Badacz ten rozróżnia zatem systemy tylko autopojetyczne od systemów adaptacyjnych, które poza autopojezą są zdolne do operacji sensotwórczych, czyli normatywnych. Tylko system, który jest w stanie zmieniać swoje środowisko w celu utrzymania się w istnieniu w niesprzyjających warunkach, może być nazwany „sprawcą autonomicznym". System taki reguluje swoje zachowanie zgodnie z normami, które

pochodzą od niego samego, czyli z zasad działania jego struktury organizacyjnej, a nie są nadane z zewnątrz przez innego sprawcę. Adaptacyjność, innymi słowy, to regulowanie zachowania w niestabilnych (ang. *precarious*)[5] okolicznościach, ocenianie różnych rozwiązań pod względem ich efektywności w utrzymywaniu autonomicznego istnienia. Tylko połączenie autopojetyczności z taką zdolnością gwarantuje autonomię sprawczą.

Rozwiązanie Di Paolo polegające na połączeniu w definicji autonomicznego działania warunków autopojetyczności i adaptacyjności jest podzielane przez innych enaktywistów. Wskazują oni, że wyróżniającą cechą autonomicznych podmiotów działania nie jest ich autopojetyczna natura, którą traktują jako warunek konieczny i oczywisty, lecz sensomotoryczna aktywność w środowisku, z której wyłania się adaptacyjne zachowanie (Thompson, 2007; Barandiaran, 2017). Myśl tę podsumowali Xabier Barandiaran, Ezequiel Di Paolo i Marieke Rohde, proponując trzy warunki autonomicznej sprawczości. Po pierwsze, system powinien ustanawiać siebie jako odrębne indywiduum, po drugie, powinien być źródłem swojej własnej aktywności w środowisku, a po trzecie, powinien on regulować swoją aktywność zgodnie z konkretnymi normami, których źródłem jest on sam. Autorzy nazwali je kolejno warunkami indywidualizacji, asymetrii i normatywności. (Barandiaran, Di Paolo i Rohde, 2009). Każdy z tych warunków rodzi szereg problemów stanowiących temat ożywionej dyskusji wśród enaktywistów i ich krytyków, po ich spełnieniu jednak, w jakiekolwiek postaci się je przyjmie, system zyskuje status autonomicznego sprawcy bez konieczności operowania językiem pojęciowym, posiadania postaw propozycjonalnych i refleksyjnej samoświadomości. Aby być podmiotem autonomicznego działania, wystarczy stanowić strukturę sprzężonych ze sobą procesów, zdolną do adaptacyjnego regulowania interakcjami ze środowiskiem zgodnie z normami określonymi przez swoje własne warunki istnienia. Normatywność systemu odnosi się do celów i zasad jego organizacji, które określają czym jest dla niego dobre, a czym złe funkcjonowanie. Wyłania się ona, co jest godne podkreślenia, z jego sensotwórczej interakcji ze środowiskiem wyznaczającej to, co jest dla niego znaczące. Wszelkie procesy poznawcze są zatem relacyjne, nie mają konkretnej lokalizacji i z tego względu enaktywiści

[5] Warunek niestabilności (ang. *precariousness*) jest zdefiniowany przez enaktywstów następująco: *in the absence of the enabling relations established by the operationally closed network, a process belonging to the network will stop or run down [...] a precarious, operationally closed system is inherently restless, and in order to sustain itself despite its intrinsic tendencies towards internal imbalance, it requires energy, matter, and relations with the outside world. [...] Hence, the system is not only self-enabling, but also shows spontaneity in its interactions due to a constitutive need to constantly "buy time" against the negative tendencies of its own parts* (Di Paolo i Thomson, 2014, s. 72).

unikają przeważnie udziału w debacie internalizmu z eksternalizmem dotyczącej tego, gdzie przebiegają procesy poznawcze (Di Paolo, 2009)[6].

Zwolennicy definiowania działania w terminach stanów umysłowych zakładających refleksyjność podmiotu jasno oddzielają jego poziom organiczny od poziomu osobowego. Zdaniem Davidsona, działanie (ang. *agency*) można wyjaśnić tylko na tym drugim poziomie, tam, gdzie pojawiają się intencje. Systemy, które nie posiadają stanów propozycjonalnych, takie jak organizm na poziomie opisu biologicznego lub sztuczny system, wykonują pewne aktywności, lecz nie stanowią sprawców działania. Enaktywiści tymczasem dopuszczają możliwość przypisania sprawczości (również w sensie ang. *agency*) wszelkim systemom wykonującym zadania poznawcze, jeśli spełniają one warunki autonomii sensomotorycznej. Być może jest to tylko kwestia werbalna, którą można rozwiązać przez wprowadzenie podziału na pojęcie sprawczości autopojetycznej i sprawczości refleksyjnej. Badacze sprawczości refleksyjnej, tacy jak Davidson, najpewniej nie zaprzeczyliby temu, że systemy autopojetyczne są autonomiczne w sensie biologicznym, wyjaśnianym przez enaktywistów. Warto jednak zwrócić uwagę na eksplanacyjną przewagę perspektywy enaktywistycznej w kontekście pytania, skąd biorą się owe wyrafinowane stany umysłowe decydujące o refleksyjności, innymi słowy, jakie minimalne warunki muszą zostać spełnione, aby system poznawczy je wytworzył. Co ciekawe, również system grupowy może być autonomicznym sprawcą w omawianym znaczeniu. Problematyka takich systemów jest analizowana w ramach enaktywistycznej teorii poznania społecznego, której przedstawiciele skupiają się na określaniu społecznych interakcji sensotwórczych pomiędzy jednostkami, do czego powrócę w ostatniej części tej pracy (De Jaegher i Di Paolo, 2007; Di Paolo i Thompson, 2014). Autonomicznymi sprawcami w znaczeniu autonomii sensomotorycznej nie muszą być zatem wyłącznie systemy żywe w biologicznym sensie. Warunki wyznaczone przez enaktywistów mogą być realizowane przez systemy, które nie są organizmami, czyli teoretycznie również przez systemy sztucznej inteligencji.

Najważniejszym warunkiem autonomii, podkreślanym przez enaktywistów, jest sensomotoryczne, cyrkularne sprzężenie systemu z jego środowiskiem, w wyniku którego wyłania się funkcjonalnie określony wzorzec neurodynamiczny, stanowiący jego strukturę organizacyjną. System poznający nie może istnieć bez tej interakcji (enakcji, ang. *enaction*), stąd teza o jego sensomotorycznej konstytucji (Barandiaran, 2017). Łączy się ona z niechęcią enaktywistów wobec stosowania kategorii reprezentacji w takiej postaci, w jakiej była ona pierwotnie wprowadzona w kognitywistyce obliczeniowej. Mianowicie podmiot poznający nie jest pasywnym odbiorcą informacji ze świata, prze-

[6] Należy jednak przy tym zaznaczyć, że enaktywiści silnie opowiadają się po stronie koncepcji umysłu ucieleśnionego w debacie ucieleśnienie/funkcjonalizm.

kształcanym następnie w jego umyśle w wewnętrzne obrazy, którym nadawane jest znaczenie. Tworzenie adekwatnych reprezentacji rzeczywistości nie jest głównym celem organizmu, który próbuje utrzymać swoje istnienie w niesprzyjających warunkach. Celem tym jest wytwarzanie znaczenia w procesie sensomotorycznej interakcji z elementami tej rzeczywistości i dążenie do zdobycia tego, czemu to znaczenie (wartość) zostało nadane.

Wyróżniającą cechą autonomicznych podmiotów działania, którą warto podkreślić przed próbą zastosowania tego pojęcia do analizy sprawczości sztucznych systemów, jest możliwość dokonywania selekcji pomiędzy alternatywami oraz realizacji jednej z nich: efektywnej lub błędnej w kontekście celu adaptacyjnego. Sprawczość pojawia się zatem tylko w tych systemach, które posiadają możliwość wyboru, co przypomina warunki stawiane jej przez zwolenników koncepcji refleksyjnych. „Wybór" jest jednak słowem mylącym w kontekście sprawczości systemów pozaludzkich, dlatego w poprzednim zdaniu używałam słowa „selekcja", które nie ma tak silnych konotacji ze świadomością refleksyjną. Działanie, według enaktywistów, w przeciwieństwie do ruchu wywołanego przez zdeterminowany prawami fizyki ciąg zdarzeń, jest skutkiem selekcji dokonywanej przez system pomiędzy różnymi możliwymi fizycznie ruchami i kierowanej jego wewnętrzną normatywnością; selekcji, która w przypadku człowieka może łączyć się ze świadomym wyborem, a przynajmniej jego poczuciem.

Kończąc to krótkie wprowadzenie do koncepcji autonomii sensomotorycznej, warto również zwrócić uwagę na problem związany z pierwszym warunkiem sprawczości wyróżnionym przez Barandiarana, Di Paolo i Rohde, który wskazuje na cechę samotworzenia się systemu, innymi słowy, samodzielnego wyodrębniania się go spośród innych bytów w świecie (ang. *the condition of individuality*) (Barandiaran, Di Paolo i Rohde, 2009). Odnosi się on do zagadnienia granic systemu poznawczego, które przez zewnętrznych obserwatorów mogą być wyznaczane dość arbitralnie w zależności od tego, jakie zadanie poznawcze jest rozpatrywane. Jak zatem określić faktyczne granice autonomicznego sprawcy i odróżnić je od tych, które nadaje systemowi obserwator, stosując na przykład strategię nastawienia intencjonalnego (Dennett, 1987)? Ontologiczne założenie leżące u postaw enaktywizmu głosi, że system, który tylko zachowuje się jak autonomiczny dla zewnętrznego obserwatora, lecz nie posiada odpowiedniego wewnętrznego mechanizmu decydującego o jego autonomii, nie stanowi autentycznego sprawcy. Problem ten jest szczególnie intrygujący w odniesieniu do sztucznych systemów, które nie posiadają, przynajmniej na dzień dzisiejszy, „rzeczywistych", samowytworzonych granic. Zgodnie z omawianą koncepcją autonomicznej sprawczości, obecnie istniejące systemy sztucznej inteligencji nie definiują się same jako odrębne byty i nie wytwarzają same swojej wewnętrznej normatywności. Programista może stworzyć odpowiednią strukturę sztucznych sieci neuronowych, która – wcho-

dząc w interakcje z elementami zaprojektowanego środowiska – uczy się zachowania realizującego ustalone przez człowieka normy. System taki nie stanowi jednak autonomicznego sensomotorycznego sprawcy.

W kolejnej części tekstu wskażę na główne przeszkody leżące na drodze ku realizacji projektu stworzenia sztucznego podmiotu autonomicznego działania oraz na próby ich ominięcia. Sukces tego przedsięwzięcia może przyczynić się do rozwiania dystopijnych wizji zagłady gatunku ludzkiego przez superinteligentą, lecz pozbawioną empatii technologię i wytworzenia nowego obrazu przyszłości: współpracy człowieka z autentycznie empatyczną sztuczną inteligencją, która podziela i przez to rozumie ludzkie potrzeby i cele.

4. Sztuczny system poznawczy jako autonomiczny sensomotoryczny sprawca

Robotycy i programiści tworzą obecnie systemy, które rozwiązują zadania poznawcze w sposób coraz bardziej niezależny od człowieka. Z tego powodu analizy pojęcia autonomicznej sprawczości, które mogłyby być zastosowane do takich systemów, zaczynają zajmować uwagę coraz szerszej grupy filozofów, w tym etyków. Jak wskazałam w poprzedniej części artykułu, definicje działania, które odwołują się do refleksyjności podmiotu i jego kompetencji językowych, nie są odpowiednim punktem startu na drodze ku stworzeniu sztucznego inteligentnego sprawcy, którego działania odpowiadałyby ludzkim wartościom i intencjom. Refleksyjność może pojawić się po osiągnięciu przez system autonomiczny pewnego stopnia złożoności, jak to miało miejsce w przypadku organizmów, w pierwszej jednak kolejności należy zadbać o to, aby stworzyć dla niej odpowiednie podstawy, czyli spełnić warunki minimalnej autonomii. Aby osiągnąć ten cel, należy zaprojektować system, który będzie zdolny do samorozwoju przez uczenie się, czyli taki, który będzie posiadał abstrakcyjną przestrzeń do samodzielnego wytwarzania się i samoutrzymania w istnieniu. W tym celu projektanci sztucznej inteligencji tworzą algorytmy, które, bazując na ograniczonych danych i ubogich warunkach początkowych, rozwijają się do złożonych procesów symulujących na przykład takie naturalne zjawiska jak system immunologiczny lub sieć neuronowa dokonująca kategoryzacji obiektów. Programy takie modelują rozwój naturalnych autonomicznych systemów, które wykazują się adaptacyjnością w odpowiedzi na zmieniające się warunki środowiskowe. Czy stanowią one jednak sprawców na podobieństwo biologicznych systemów? Jeśli nie, co stanowi różnicę i czy można ją będzie wyeliminować na gruncie rozwijającej się technologii? Odpowiedź na te pytania wymaga ogromu pracy na gruncie symulacji komputerowych i modeli robotycznych, jak również pogłębionych badań biochemicznych i biotechnologicznych. Nie jest ona również możliwa bez analiz filozoficznych, a takie zapewnia między innymi omawiany w tej pracy enaktywizm.

Założeniem projektu stworzenia sztucznego sensomotorycznego sprawcy jest teza, że autopojeza może wyłonić się z innych procesów niż te, które są odpowiedzialne za powstanie biologicznego życia. System taki byłby funkcjonalnie nieodróżnialny od autopojetycznego organizmu, chociaż nie byłby żywy w biologicznym sensie. Czy możliwe jest jednak stworzenie sztucznego systemu, któremu pewne własności środowiska prezentowałyby się jako znaczące z jego własnej wewnętrznej perspektywy, a nie z perspektywy projektanta? Pytanie to przybiera postać problemu znaczenia w sztucznych systemach inteligentnych i jest przedmiotem ożywionej dyskusji filozoficznej.

Jednym z badaczy sceptycznie nastawionych do sensotwórczych możliwości sztucznej inteligencji jest Hubert Dryefus. Jego zdaniem, w celu osiągnięcia sukcesu na tym gruncie musielibyśmy zaimplementować systemowi model ciała, które jest zakorzenione w środowisku w sposób podobny do ciała ludzkiego. Model taki wymagałby szczegółowego opisu ludzkiego ciała i ludzkich motywacji, tylko wtedy bowiem elementy świata byłyby dla niego znaczące w podobny sposób jak dla człowieka, co stanowi konieczny warunek uczenia się inteligentnego zachowania w rzeczywistym świecie (Dryefus, 2007). Krytycy takiego podejścia zauważają jednak, że jest ono zbyt wymagające. Zamiast tworzyć szczegółowy opis działania żywych organizmów w całej ich złożoności, wystarczy określić warunki powstania najprostszego systemu, który dysponowałby zdolnościami sensotwórczymi (Di Paolo i Iizuka, 2008; Morse i Ziemke, 2008). Warunek zakorzenienia w środowisku, z którym system wchodzi w interakcje, jest warunkiem koniecznym, lecz niewystarczającym wewnętrznej normatywności. Należy dodać do niego warunek, który sprawi, że systemowi będzie na czymś „zależeć", że będzie on „chciał" zdobyć to, co zapewni mu dalsze istnienie, które jest jego własnym, a nie zadanym z zewnątrz, celem (Di Paolo, 2003). Nie wystarczy zatem zaimplementowanie systemowi hierarchii wartości, z których najwyższą pozycję zajmuje wartość przetrwania, wciąż bowiem będzie to norma zadana z zewnątrz.

Różnicę pomiędzy żywymi a sztucznymi systemami poznawczymi w ciekawy sposób przedstawiają Tom Froese i Tom Ziemke (2009). Tkwi ona, według nich, w sposobie istnienia tych obiektów. Podczas gdy sztuczne systemy mogą istnieć nawet wtedy, gdy nie wykonują żadnego działania (np. wyłączony robot), żywe organizmy muszą angażować się w procesy adaptacyjne, aby kontynuować swoje istnienie. Jest to zatem różnica pomiędzy „istnieniem poprzez istnienie" (ang. *being by being*) a „istnieniem poprzez działanie" (ang. *being by doing*). Wewnętrzna normatywność jest więc ugruntowana w niestabilności systemów podlegających metabolizmowi. Chociaż wydawać by się mogło, że wniosek ten zamyka drogę wszelkim próbom stworzenia sztucznego autonomicznego sprawcy, który nie jest biologicznym organizmem, koncepcja systemu autopojetycznego Vareli i Maturany pozwala ominąć tę przeszkodę. Na jej gruncie bowiem metabolizm może być zinterpretowany jako stanowiący jedną z możliwych – fizykochemiczną – realizację bardziej ogólnej zasady domknięcia orga-

nizacyjnego (ang. *organizational closure*), która odpowiada za autonomię systemu (Varela, 1979; Maturana i Varela, 1980). Zatem pomimo że autopojetyczność jest cechą systemu żywego, autonomia może cechować system, który nie przeprowadza metabolizmu (Thompson, 2007).

Opierając się na tego typu rozważaniach, Froese i Ziemke proponują dwie enaktywistyczne zasady projektowania autonomicznej sztucznej inteligencji. Pierwsza z nich głosi, że system musi być zdolny do wytwarzania swojej własnej systemowej jedności (ang. *systemic identity*) na pewnym poziomie opisu. Oznacza to, że posiada on wewnętrzną normatywność wyznaczającą jego interakcje z elementami środowiska, które są dla niego znaczące, gdyż umożliwiają mu utrzymanie się w istnieniu w niestabilnych warunkach. Druga zasada, wprowadzona przez omawianych badaczy, nakłada na system warunek polegający na konieczności aktywnego, adaptacyjnego regulowania ciągłej sensomotorycznej aktywności w relacji do różnych ograniczeń wewnętrznych i środowiskowych (Froese i Ziemke, 2009). Owe dwie zasady odpowiadają omówionej w poprzedniej części artykułu definicji autonomii sensomotorycznej zaproponowanej przed Di Paolo. Przypomnę, że badacz ten łączy warunek autopojetyczności z warunkiem adaptacyjności. Pierwszy z nich, jak wskazałam powyżej, może – pod postacią warunku domknięcia organizacyjnego, umożliwiającego systemowi samodzielne utrzymanie się w istnieniu i regenerację – zostać rozszerzony na systemy, które nie przejawiają biologicznego życia.

Wskazane powyżej enaktywistyczne warunki autonomicznego działania są niezwykle wymagające w kontekście obecnie tworzonych programów sztucznej inteligencji. Sztuczny system, który by je spełniał, musiałby istnieć w sposób przejawiany obecnie tylko przez organizmy, mianowicie musiałby „istnieć poprzez działanie", a nie tylko „poprzez istnienie". Znaczy to, że jego istnienie byłoby niestabilne i zależne od ciągłego procesu samoodnowy i przystosowywania się do warunków środowiskowych; wyłączenie go byłoby, jednym słowem, analogiczne do biologicznej śmierci żywego systemu. Normatywność sztucznego sprawcy musiałaby wyłaniać się z jego interakcji z niestabilnym środowiskiem, tak aby cele jego działania były autentycznie jego, a nie zadane przez programistę. Jednym z najbardziej obiecujących pod tym względem systemów sztucznej inteligencji jest ewolucyjny system obliczeniowy EMOTIONAL (*EMbodied Open-ended evoluTIONary ALgorithm*) zaproponowany przez Yuriego L.B. Noguerię i jego współpracowników, który w zamierzeniu twórców miał przejawiać wewnętrzną autonomię (Nogueira i in., 2019). Projektanci tego systemu nie wprowadzili do niego żadnego innego celu poza przetrwaniem, po to, aby to on sam wytwarzał własne reguły interakcji z elementami jego środowiska. Można zatem powiedzieć, że system ten jest adaptacyjny, uczy się bowiem, w interakcji ze swoim otoczeniem, jak regulować własne zachowanie i co jest dla niego znaczące. Nie spełnia on jednak pierwszego warunku Froese'go i Ziemke'go, który mówi o samodzielnym wytwarzaniu przez system swojej własnej jedności (ang. *identity*), czyli samodzielnym wyodrębnianiu się

spośród innych obiektów w świecie. Wciąż to projektant wyróżnia system z otoczenia i definiuje go jako odrębny obiekt, nawet jeśli czyni to w najbardziej minimalnej formie.

Współczesnym przejawom sztucznej inteligencji brakuje zatem autopojetyczności w sensie metabolicznej samoorganizacji. Jak już wspomniałam, spełnienie tego warunku nie jest konieczne dla uzyskania statusu autonomicznego sprawcy, gdyż mechanizm decydujący o domknięciu organizacyjnym, czyli o samodzielnym wyróżnianiu się systemu z otoczenia i określaniu swoich granic, może być realizowany poprzez inne fizyczne procesy niż te, które są odpowiedzialne za powstanie biologicznego życia. W sukces takiego przedsięwzięcia wątpią między innymi Alvaro Moreno i Arantza Etxeberria. W artykule „Agency in Natural and Artificial Systems" (2005) wskazują oni, że system spełniający warunki sensomotorycznej autonomii musi posiadać wewnętrzny punkt widzenia, perspektywę, z której pewnym elementom otoczenia nadawane jest znaczenie, a innym nie. Aby wykazywać się autentyczną sprawczością, sztuczny system musiałby wykonywać swoje działanie nie dlatego, że został zaprogramowany do pewnego celu, lecz w pierwszej kolejności dlatego, żeby utrzymać się w istnieniu. To ta wewnętrzna motywacja, podstawowa chęć „najpierw przetrwać" rozróżnia biologiczne systemy, które „istnieją poprzez działanie", od sztucznych, które istnieją nawet wtedy, gdy nie rozwiązują żadnych zadań. Moreno i Etxeberria wskazują, że autonomia oparta na procesach samowytwarzania i samoregeneracji jest aktualnie realizowana tylko w żywych organizmach przez wysoce złożoną organizację chemiczną; odpowiednia przemiana energii zapewniająca domknięcie organizacyjne może być bowiem osiągnięta tylko poprzez metabolizm, na który składają się konkretne reakcje chemiczne (Moreno i Etxeberria, 2005). Sukces projektu stworzenia autonomicznego sztucznego systemu zależy zatem od tego, czy możliwe jest (lub będzie) osiągnięcie materialnie niestabilnego systemu, którego procesy adaptacyjne i sensotwórcze nie są skutkiem metabolizmu, lecz wyłaniają się z innego rodzaju zasad organizacyjnych. Omawiani autorzy nie mają zbyt wielkich nadziei na powodzenie tego przedsięwzięcia. Niektóre programy sztucznej inteligencji wykazują, co prawda, zdolności adaptacyjne w zaprojektowanym środowisku, lecz wciąż brakuje im podstawowej autopojetyczności, polegającej na samodzielnym wyróżnianiu się z otoczenia jako odrębnego materialnego bytu, dla którego działanie jest konieczne dla istnienia.

Chociaż stworzenie sztucznego systemu cechującego się autonomią sensomotoryczną stanowi ogromne wyzwanie, wciąż podejmowane są próby motywowane nadzieją na sukces (Woolford i Egbert, 2022; Xie i Jin, 1918). W ostatniej części tej pracy wskażę, z czym konkretnie owe nadzieje się wiążą i po co właściwie podejmować próby tworzenia systemu, który jest równie wrażliwy i niestabilny jak żywy organizm, podczas gdy mamy już możliwości technologiczne do tworzenia systemów dużo lepiej radzących sobie z niektórymi zadaniami

poznawczymi od człowieka. Czy nie lepiej skupić się na ulepszaniu konkretnych programów sztucznej inteligencji zaprojektowanych do osiągania precyzyjnych celów, zamiast usiłować stworzyć system, który „wytwarza się sam" i początkowo nie ma żadnego innego celu poza przetrwaniem, a kolejne cele wyznacza sobie sam? Nadzieje związane z takimi systemami wiążą się z zadaniami poznawczymi, które wymagają zrozumienia ludzkich wartości, potrzeb i oczekiwań, i dają efekty, które są przez człowieka oceniane jako dobre lub złe pod względem moralnym. Sztuczny system, który przejawiałby takie zrozumienie, musiałby kierować się wartościami zbliżonymi do ludzkich. To, jak osiągnąć ten cel, jest tematem rozważań prowadzonych w ramach problemu uzgodnienia wartości, który omówię w ostatniej części tej pracy.

5. Zastosowania enaktywistycznej koncepcji autonomii sprawczej w dyskusjach nad etyką sztucznej inteligencji i podmiotowością systemów grupowych

Czym różni się inteligencja żywych organizmów od sztucznych systemów rozwiązujących zadania poznawcze? Najogólniej, inteligencję można określić jako zdolność do właściwego używania wiedzy w celu poprawiania dokładności przewidywań dotyczących przyszłych zdarzeń i planowania działania tak, aby utrzymać się w istnieniu w jak najlepszej kondycji (Dennett, 1984). Tak zdefiniowana inteligencja odnosi się jednak tylko do tych systemów, które są niestabilne, ich istnienie zależy od ich działania i mają za cel „przede wszystkim przetrwanie". Sztuczne systemy nie spełniają obecnie tych warunków. Nie przejawiają one inteligencji, od której zależy ich istnienie, lecz taką, która umożliwia rozwiązanie konkretnych zadań poznawczych zależnych od projektanta. Wiele zadań relegowanych na sztuczne systemy wymaga jednak zrozumienia ludzkich intencji, a co za tym idzie, wartości i oczekiwań, które trudno jest zaprogramować systemowi wprost. Nawet jeśli system posiada dostęp do ogromnej bazy danych, musi być w stanie wybrać spośród nich te informacje, które są w danej sytuacji istotne dla rozwiązania problemu w sposób oczekiwany przez człowieka. Jest to tak zwany problem ramy (ang. *the frame problem*), który pierwotnie został sformułowany jako wyzwanie polegające na reprezentowaniu w logice skutków danego działania bez konieczności jawnego reprezentowania dużej liczby zdarzeń, które intuicyjnie nie stanowią istotnych skutków (McCarthy i Hayes, 1969). Problem ten przedostał się jednak szybko na grunt epistemologii i filozofii sztucznej inteligencji i przybrał formę pytania, w jaki sposób wyjaśnić zdolność podejmowania decyzji przez żywy system poznawczy wyłącznie na podstawie tego, co jest istotne w bieżącej sytuacji, bez konieczności rozważania przez niego wszystkich dostępnych danych, które istotne nie są (Dennett, 1984). Wydaje się, że zdolność ta należy do istoty naturalnej inteligencji, czy można ją jednak uzyskać w systemie sztucznym? System taki musiałby ignorować wiele danych, które posiada, i operować małym zbiorem umiejętnie i szybko wybranych informacji, które są przydatne w danym momencie w dynamicznie zmieniających się okolicznościach. Takie istotne

dane musiałby on dobierać ze względu na oczekiwany przez człowieka rezultat danego działania, co wymagałoby rozpoznania intencji i wartości, które są poprzez to działanie realizowane. Celem jest zatem stworzenie sztucznego systemu, który przejawia naturalną inteligencję charakterystyczną dla autonomiicznych sprawców sensomotorycznych.

Problem ramy wiąże się ze wspomnianym już problemem uzgodnienia wartości, który wyrażony jest w pytaniu o to, jak upodobnić wartości, którymi kieruje się sztuczny system do wartości ludzkich, po to, aby rozwiązywał on zlecone mu zadania zgodnie z ludzkimi intencjami i oczekiwaniami, i co za tym idzie, nie stanowił dla człowieka zagrożenia (Gabriel, 2020; Christian, 2021; Cannon, 2021). Innymi słowy, nawiązując do problemu ramy: jak sprawić, aby system, spośród ogromnego zbioru danych, do których ma dostęp podczas rozwiązywania danego problemu, wybierał te informacje, które są istotne z ludzkiej perspektywy? Enaktywistyczna koncepcja sprawczości sensomotorycznej niesie ze sobą nadzieję na stworzenie systemu, który po spełnieniu warunków autonomicznego działania będzie przeprowadzał podobne procesy sensotwórcze i tym samym będzie wrażliwy na podobne wartości co żywy organizm.

Problem uzgodnienia wartości narzuca się szczególnie w dyskusji nad konsekwencjami powstania ogólnej sztucznej inteligencji, która przewyższałaby człowieka w rozwiązywaniu każdego zadania poznawczego, a nie, tak jak to jest teraz, w jednym lub kilku konkretnych zadaniach (Bostrom, 2014; Russell, 2019). W największym skrócie: zagrożenie polega na tym, że mając przed sobą jakikolwiek cel, inteligencja taka będzie starała się osiągnąć go najbardziej racjonalnymi i efektywnymi metodami, co może doprowadzić do konfliktu z ludzkimi wartościami. Michael Cannon, rozważając ten problem, wskazuje na konieczność rozróżnienia pomiędzy efektywnym rozwiązywaniem problemu przez sztuczny system a rozwiązywaniem go w odpowiedni sposób, czyli taki, który współgra z ludzkimi wartościami, jakkolwiek rozumianymi. Bez takiego rozróżnienia istnieje zagrożenie, że sztuczna inteligencja, w celu na przykład rozwiązania problemu wojen religijnych, podejmie decyzję o zakazie wyrażania wszelkich uczuć religijnych.

Na gruncie etyki maszynowej i filozofii moralności trwa żywa dyskusja dotycząca możliwości stworzenia moralnego sztucznego systemu (Anderson i Anderson, 2011; Behdadi i Munthe, 2020; Coeckelbergh, 2009; De Poel, 2020; Hagendorff, 2020; Himma, 2009; Sparrow, 2021). Kierunek wyznaczony przez zwolenników enaktywizmu wydaje się jednym z najbardziej obiecujących, lecz jednocześnie jednym z najbardziej odległych. Ostatecznie, celem tego przedsięwzięcia jest powołanie do istnienia sztucznego refleksyjnego podmiotu działania. Aby to jednak było możliwe, najpierw musi powstać sensomotoryczny autonomiczny system w najprostszej postaci, którego głównym celem byłoby utrzymanie się w istnieniu i byłby to cel wyłaniający się z jego wewnętrznej struktury organizacyjnej, a nie zadany z zewnątrz przez innego sprawcę. System taki rozwijałby inteligencję, która nie polega wyłącznie na efektywnym

rozwiązaniu konkretnego, zadanego z zewnątrz problemu, lecz objawia się umiejętnością samodzielnego wyznaczenia celu działania. Tym właśnie różnią się obecne systemy sztucznej inteligencji od autonomicznego sprawcy, mianowicie potrafią one sprawnie, a nawet nowatorsko rozwiązywać problemy, lecz nie wykazują umiejętności ich formułowania w sposób, który byłby dla nich samych egzystencjalnie ważny. Definiowanie problemu nie zawiera się w pojęciu sztucznej inteligencji, którym dysponujemy obecnie, obejmuje ono bowiem tylko rozwiązywanie określonych wcześniej problemów w najbardziej optymalny sposób (Cannon, 2021). Trudność pojawia się jednak w momencie, gdy sztuczny system źle zinterpretuje cel, gdyż nie rozumie oczekiwań i intencji w nim zawartych. Gdyby możliwe było stworzenie sztucznej inteligencji, która potrafi określić problem, przed którym stoi i powiązać z nim odpowiednie wartości, ryzyko, że rozwiąże go niezgodnie z ludzkimi oczekiwaniami byłoby znacznie ograniczone.

Michael Cannonn (2021) podsumowuje enaktywistyczną propozycję rozwiązania problemu uzgodnienia wartości następująco: zakłada ona, bazując na sensomotorycznej koncepcji autonomii, że to, co jest wartościowe i istotne dla danego systemu w danej sytuacji poznawczej, zależy od tego, jakim jest on bytem pod względem ontologicznym. Jeśli systemy są ontologicznie do siebie podobne, podzielają one podobne wartości. Zatem jeśli człowiek i sztuczna inteligencja byliby ontologicznie do siebie podobni, podzielaliby podobne wartości. Mówiąc o podobieństwie ontologicznym, Cannon odwołuje się do określeń systemu poznającego formułowanych na gruncie enaktywizmu w postaci tez zgromadzonych pod wspólną nazwą „4Es" (ang. *embodiment, embeddedness, extendedness, and enaction*). Można je jednak również przedstawić w postaci warunków autonomii sensomotorycznej: autopojetyczności (lub jej niebiologicznego odpowiednika) i adaptacyjności. Wszystkie one wytwarzają „dziedzinę istotności" (ang. *domain of relevance*), która powinna stanowić zasadę projektowania sztucznych autonomicznych sprawców w przedsięwzięciach, których celem jest uzgodnienie wartości sztucznej inteligencji z wartościami ludzkimi. Należy oczywiście zauważyć, co robi Cannon, że ontologiczne podobieństwo nie gwarantuje dokładnego uzgodnienia wartości i związanego z nim bezpieczeństwa, co unaocznia wizja zagłady nuklearnej, którą ludzie sami mogą na siebie sprowadzić. Sprzyja jednak ono temu, że podobne problemy będą w podobny sposób istotne dla sztucznej inteligencji, co może wpłynąć na jej rozumienie intencji i oczekiwań ludzi, a to, czy postąpi ona zgodnie z nimi, stanowi kolejny problem. W pierwszej kolejności należy projektować sztuczne systemy autonomiczne w taki sposób, aby miały te same ograniczenia ontologiczne co ludzie, gdyż definiują one dziedzinę tego, co jest dla systemu istotne. Takie „uzgodnienie ontologiczne" (ang. *ontological alignment*) daje szanse na to, że dwa systemy (człowiek i SI) będą posiadały tę samą definicję problemu dotyczącego na przykład ochrony środowiska, bezrobocia czy wyboru metod leczenia pacjentów, a w związku z tym podobne metody jego rozwiązania, oparte na podobnych wartościach. Najważniejsza róż-

nica pomiędzy takimi projektami a tymi, które są realizowane, leży obecnie w sposobie określania problemu i odpowiednich metod jego rozwiązania. To, co istotne dla obecnych programów sztucznej inteligencji, jest zdefiniowane przez problem sformułowany przez programistę dla programów, które byłyby autonomiczne w sensie enaktywistycznym, a to, co istotne, byłoby zdefiniowane przez sam program.

Pomimo prób stworzenia sztucznego systemu, który charakteryzowałby się sensomotoryczną sprawczością, cel ten jest wciąż poza zasięgiem człowieka. Rozważania na ten temat są jednak cenne, gdyż rzucają światło na różnice pomiędzy autentycznymi sprawcami a sztucznymi systemami. Świadomość owych różnic pozwala zrozumieć ograniczenia sztucznej inteligencji i związane z nimi zagrożenia. Rozwiązując jakiekolwiek zadanie, sztuczny system może wybrać metodę, która jest dla człowieka nieprzewidywalna, gdyż baza danych, z której korzysta, szukając najbardziej efektywnego rozwiązania, jest ogromna. Stanowisko enaktywistyczne wobec sprawczości stanowi teoretyczną próbę rozwiązania tego problemu, wskazuje mianowicie, z jakich minimalnych warunków wyłoniła się ludzka sprawczość i refleksyjność oraz jaki kierunek należy obrać, aby stworzyć sztuczny system, który czegoś chce i któremu na czymś zależy.

Kończąc rozważania dotyczące zastosowania sensomotorycznej koncepcji autonomii do analizy podmiotowości systemów, które nie przejawiają jako takie życia, warto wspomnieć ponownie o problematyce sprawczości systemów grupowych, złożonych z więcej niż dwóch ludzi. Traktowanie takich systemów jako podmiotów refleksyjnych budzi wątpliwość, gdyż wiąże się z przypisywaniem im stanów umysłowych, które nie są stanami poszczególnych członków grupy, tylko należą do niej, jako odrębnego, emergentnego podmiotu. Najbardziej nieintuicyjne wydaje się przypisywanie grupom wewnętrznego punktu widzenia, z którego system refleksyjny rozpoznaje siebie jako podmiot swych działań posiadający kontrolę nad swoimi wyborami. Koncepcja autonomii sensomotorycznej wydaje się lepszym narzędziem do analizy sprawczości takich systemów, gdyż zakłada mniej wymagające warunki. Jeśli przyjmiemy, że cechy domknięcia organizacyjnego, samotworzenia i adaptacyjności mogą zostać zrealizowane w systemach niebiologicznych, droga do analizy systemów grupowych jako autonomicznych sprawców zostaje otwarta, podobnie jak to ma miejsce w odniesieniu do systemów sztucznych. Najważniejsza różnica pomiędzy systemami biologicznymi a grupowymi dotyczy ich komponentów, którymi w przypadku grup są refleksyjni sprawcy. Interakcje między nimi nie polegają zatem wyłącznie na relacjach przyczynowych pomiędzy fizycznymi elementami, lecz również na komunikacji językowej rozumianej jako relacja pomiędzy znaczeniami. Z tego względu grupa jako autonomiczny sprawca powstaje w wyniku samoorganizacji między jednostkowymi autonomicznymi sprawcami, którzy mogą podejmować decyzję o byciu jej częścią bądź nie, o swojej

roli w niej, a także o ogólnej kondycji grupy jako podmiotu działania (Goldspink i Kay, 2003).

System grupowy wykazuje organizacyjne domknięcie, gdyż jego jedność, trwałość i adaptacyjność jest zdeterminowana przez wewnętrzne relacje pomiędzy jego członkami. Grupa posiada również wewnętrzną normatywność, gdyż cele kierujące jej działaniem nie są zadane z zewnątrz przez innego sprawcę, lecz wyłaniają się z jej struktury organizacyjnej i procesów sensotwórczych, które zwrotnie wpływają na relacje pomiędzy jej członkami. Enaktywistyczna koncepcja sprawczości umożliwia zatem wyjaśnienie działania systemów grupowych jako autentycznych sprawców wykazujących cechy charakterystyczne dla organizmów, takie jak autonomia organizacyjna i adaptacyjność. Taką strategię przyjmuje między innymi Niklas Luhmann (1995, który wskazuje na komunikację językową jako na relację tworzącą strukturę organizacyjną grupy odpowiedzialną za jej autonomię sprawczą. Jego analizy potwierdzają tezę, że procesy samotworzenia i samoutrzymania nie muszą zachodzić tylko w żyjących systemach, gdyż te, które nie przejawiają własności życia, również są zdolne do procesów sensotwórczych kierujących ich aktywnością.

Na rzecz traktowania grup jako pełnoprawnych podmiotów działania argumentuje również inny znany badacz tego tematu Dave Elder-Vass (2010). W jego przekonaniu, grupa jest samodzielnym, emergentnym bytem przejawiającym własną moc sprawczą nieredukowalną do mocy przyczynowych jej członków. Jego analizy w dużym stopniu współgrają z enaktywistycznymi wyjaśnieniami wewnętrznej normatywności systemów autonomicznych, gdyż określają grupę jako system, którego zachowanie determinuje jego własna struktura organizacyjna, określająca i zwrotnie określana przez normy wyznaczające warunki jej istnienia. Mechanizm ten na poziomie grupy stanowi źródło wszystkich decyzji, przekonań i wartości, zatem tak jak w przypadku wszystkich autonomicznych sprawców, normy jej działania nie pochodzą z zewnątrz, lecz są efektem i zarazem przyczyną jej działania.

Analizą sprawczości systemów grupowych zajmują się również badacze rozwijający enaktywistyczną teorię poznania społecznego (De Jaegher, Di Paolo 2007, Di Paolo, Thompson 2014). Definiują oni mianowicie relacje społeczne w terminach wzajemnego tworzenia sensu (ang. *joint sens-making*), który przejawia się w wartościach i celach grupy. Interakcja społeczna pomiędzy członkami grupy tworzy i utrzymuje jej jedność jako osobnego systemu, który, tak jak inni autonomiczni sprawcy, wykazuje niestabilność, domknięcie organizacyjne i adaptacyjność związaną z możliwością przeprowadzania procesów sensotwórczych[7].

[7] Więcej na temat wykorzystania enaktywistycznej koncepcji sprawczości w analizach dotyczących podmiotowości systemów grupowych można przeczytać w moim artykule: Tomczyk, 2024.

Rozważania wskazane w tej części artykułu są przykładem dyskusji, w których przedstawiciele stanowiska enaktywistycznego wobec sprawczości wyrażają ważny głos. Zarówno w dyskusjach dotyczących rozwijającej się sprawczości sztucznej inteligencji, jak i w tych, które analizują działania systemów grupowych, enaktywistyczna koncepcja autonomii przejawia eksplanacyjną przewagę nad stanowiskami definiującymi sprawczość w terminach stanów umysłowych zakładających refleksyjność podmiotu. Odwołując się do warunków autonomii sensomotorycznej, można wskazać źródła problemu z uzgodnieniem wartości sztucznej inteligencji i człowieka i sformułować teoretyczny pomysł na jego rozwiązanie (Cannon, 2021). Warto zauważyć, że wyjaśnianie działania systemów grupowych w terminach enaktywistycznych jest łatwiejsze i bardziej intuicyjne niż projektowanie w odwołaniu do tych terminów systemów sztucznych. Mimo że grupy jako całości również nie przejawiają życia, składają się jednak z żywych, autonomicznych sprawców, których świadome działania zapewniają spełnienie warunków autonomii. Stanowi to dużą przewagę systemów grupowych nad systemami sztucznymi, jeśli chodzi o możliwość wyjaśniania ich działania w kategoriach enaktywistycznie rozumianej sprawczości. W kontekście systemów sztucznych enaktywizm pozwala zrozumieć, że sprawczość nie zasadza się wyłącznie na rozwiązywaniu problemów, nie wymaga ona jednak również refleksyjności, przejawia się natomiast w aktywności sensotwórczej zmierzającej do realizacji wartości ważnych dla samego systemu, które wyłaniają się z jego samowytworzonej złożoności organizacyjnej. Tylko taki system miałby podstawy do tego, aby zrozumieć wartości kryjące się za postawionym przez człowieka zadaniem. W tę stronę zatem idą próby stworzenia systemu, którego efekty działania są dla człowieka moralnie znaczące.

6. Zakończenie

W pracy tej przywołałam dwie koncepcje sprawczości i wskazałam na różnice pomiędzy nimi w kontekście przypisywania autonomicznego działania systemom pozaludzkim. Pierwszą z nich nazwałam „refleksyjną", gdyż w jej ramach autonomiczne działanie definiowane jest w terminach postaw propozycjonalnych, które wymagają od podmiotu refleksyjności rozumianej jako dostęp do treści własnych stanów umysłowych, namysł nad ich pochodzeniem i racjonalnością, a także jako doświadczenie własnej podmio-towości. Jest to maksymalistyczne rozumienie sprawczości, którą przypisać można obecnie tylko człowiekowi. Koncepcja ta jest użyteczna i niezbędna przy ocenie odpowiedzialności, jaką podmiot ponosi za wykonane działanie, a która wymaga przypisania mu zasługi lub winy. Tylko podmiot refleksyjny może być w ten sposób oceniany. Druga koncepcja sprawczości, którą przywołałam, nazwana „sensomotoryczną", powstała na gruncie enaktywizmu i obejmuje dużo szerszy zakres systemów poznawczych niż pierwsza z nich. W jej ramach możliwe jest

przypisanie autonomii sprawczej systemom, które nie spełniają warunków odpowiedzialności moralnej i prawnej. Wykazują one jednak cechy minimalnej autonomii polegającej na samodzielnym konstytuowaniu swojego istnienia i adaptowaniu się do niestabilnych warunków środowiskowych.

Koncepcja autonomii sensomotorycznej pozwala na odróżnienie systemów autentycznie normatywnych, których istnienie uzależnione jest od ich działania, od systemów, których normatywność jest pochodna, czyli zadana przez innego sprawcę. Pierwsze z nich są systemami, które po uzyskaniu odpowiedniej złożoności organizacyjnej, mogą stać się refleksyjne, przeprowadzają bowiem procesy sensotwórcze, z których wyłania się, wraz ze wzrostem ich złożoności, coraz bardziej rozwinięta autentyczna normatywność. Enaktywistyczna koncepcja sprawczości ma zatem tę przewagę, że określa bazowe warunki wszelkiego działania od najprostszego po najbardziej wyrafinowane. Pozwala to na zastosowanie jej do analiz problemów filozoficznych, w tym etycznych, związanych z działaniem takich systemów jak artefakty lub grupy, które nie spełniają warunków refleksyjnej sprawczości. Efekty działania takich systemów, szczególnie jeśli chodzi o sztuczną inteligencję, są w coraz większym stopniu moralnie znaczące dla człowieka. Widać to niemal w każdej dziedzinie ich zastosowań, między innymi w służbie zdrowia, transporcie, wojnie, ochronie środowiska czy kampaniach wyborczych. Systemy takie w coraz mniejszym stopniu wymagają bezpośredniej kontroli człowieka nad sposobem rozwiązania danego zadania i być może w niedalekiej przyszłości kontrola taka, ze względu na ilość dostępnych danych i szybkość ich przetwarzania, nie będzie możliwa. Niezbędne jest zatem przemyślenie tego, z jakimi inteligentnymi sztucznymi sprawcami chcemy mieć do czynienia i jak to osiągnąć.

Bibliografia

Anderson, M. i Anderson, S. L. (red.). (2011). *Machine Ethics*. Cambridge: Cambridge University Press.

Barandiaran X. E., Di Paolo i E., Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior*, 17(5), 367–386.

Barandiaran, X. E. (2017). Autonomy and enactivism: Towards a theory of sensorimotor autonomous agency. *Topoi*, 36, 409–430.

Bayne, T. (2008). The Phenomenology of Agency. *Philosophy Compass* 1/3, 182-202 .

Beer, R. D. i Di Paolo, E. A. (2023). The theoretical foundations of enaction: Precariousness, *Biosystems*, 223, 104823.

Behdadi, D. i Munthe, C. A. (2020). Normative Approach to Artificial Moral Agency. *Minds & Machines* 30, 195–218.

Bostrom N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bratman, M. (1979). Practical Reasoning and Weakness of the Will. *Noûs* 13, 131–51.

Bremer, J. (2016). *Naukowy i potoczny obraz osoby w kognitywistyce*. Kraków: Wydawnictwo WAM.

Carter, J. A. i Palermos, S. O. (2016). The ethics of extended cognition: Is having your computer compromised a personal assault? *Journal of the American Philosophical Association, 2*(4), 542–560.

Cannon, M. (2022). An Enactive Approach to Value Alignment in Artificial Intelligence: A Matter of Relevance. W: V.C. Müller, (red.), *Philosophy and Theory of Artificial Intelligence 2021*. PTAI 2021. Studies in Applied Philosophy, Epistemology and Rational Ethics, vol 63. Springer, Cham. https://doi.org/10.1007/978-3-031-09153-7_10.

Cash, M. (2010). Extended cognition, personal responsibility, and relational autonomy. *Phenomenology and the Cognitive Sciences* 9, 645–671.

Chisholm, R. (1976). The Agent as Cause. W: M. Brand, D. Walton (red.), *Action Theory*, Dordrecht: D. Reidel Publishing Co., 199-211.

Christian, B. (2020). T*he alignment problem: Machine learning and human values*. W. W. Norton & Company.

Christman, J. (1991). Autonomy and Personal History. *Canadian Journal of Philosophy,* vol. 21, 1–24.

Christman, J. (2004). Relational autonomy, liberal individualism, and the social constitution of selves. *Philosophical Studies,* 117, 143–164.

Clark A. (2008). *Supersizing the Mind. Embodiment, Action and Cognitive Extension*. Oxford: Oxford University Press.

Clarke, R. i Reed T. (2015). Free Will and Agential Powers. *Oxford Studies in Agency and Responsibility*, 3(1), 6–33.

Coeckelbergh M. (2009). Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents. AI & Society, 24, 181–189.

Davidson, D. (1967/2002). Causal Relations. W: D. Davidson, *Essays on Actions and Events*. Oxford: Clarendon Press, 149–162.

De Jaegher, H. i Di Paolo, E. (2007). Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the cognitive sciences,* 6, 485–507.

Dennett D., (1984). Cognitive wheels: the frame problem of AI. W: C. Hookway (red.), *Minds, Machines and Evolution*. Cambridge: Cambridge University Press, 129–150.

Dennett D. (1987). *The intentional stance*. Cambridge, MA: The MIT Press.

Di Paolo, E. A. (2003). Organismically-inspired robotics: homeostatic adaptation and teleology beyond the closed sensorimotor loop. W: K. Murase, T. Asakura (red.), *Dynamical Systems Approach to Embodiment and Sociality*, Adelaide, Australia: Advanced Knowledge International, 19–42.

Di Paolo, E. A. i H. Iizuka. (2008). How (not) to model autonomous behavior. *Biosystems*, vol. 91 (2), 409–423.

Di Paolo, E. A. (2009), Extended life, *Topoi*, 28 (1), 9–21.

Di Paolo, E. i Thompson, E. (2014). The enactive approach. *The Routledge handbook of embodied cognition*, Routledge, 68–78.

Dreyfus H. L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Artificial Intelligence* 171, 1137–1160.

Dworkin, G. (1988). *The Theory and Practice of Autonomy*. Cambridge: Cambridge University Press.

Elder-Vass D. (2010). *The Causal Power Or Social Structures. Emergence, Structure and Agency*. New York: Cambridge University Press.

Farina, M. i Lavazza, A. (2022). Incorporation, Transparency and Cognitive Extension: Why the Distinction Between Embedded and Extended Might Be More Important to Ethics Than to Metaphysics. *Philosophy & Technology*, 35 (1), 1–21.

Farina, M. i Lavazza, A. (2022a). Mind embedded or extended: transhumanist and posthumanist reflections in support of the extended mind thesis. *Synthese,* 200, 507.

Fischer, J. M. i Ravizza M. (1993), *Perspectives on Moral Responsibility*, Ithaca: Cornell University Press.

Frankfurt, H. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy* , 5–20.

Froese T. i Ziemke T. (2009). Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Inteligence* vol. 173 (3–4), 466-500.

Gabriel, I. (2020). Artificial Intelligence, Values and Alignment. *Minds and Machines*, *30*(3), 411–437.

Giubilini, A. i Savulescu, J. (2018). The artificial moral advisor. The 'ideal observer' meets artificial intelligence. *Philosophy and Technology*, vol. 31, 169–188.

Goldsping Ch. i Kay, R. (2003). Organizations as self-organizing and sustaining systems: A complex and autopoietic systems perspective. *International Journal of General Systems*, 32(5), 459–474.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds & Machines* 30, 99–120.

Himma, K.E. (2009). Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? *Ethics Inf Technol,* 11, 19–29.

Lara, F. i Deckers, J. (2020). Artificial intelligence as a socratic assistant for moral enhancement. *Neuroethics*, *13*(3), 275–287.

Luhmann N. (1995). *Social Systems*, Stanford: Stanford University Press.

Maturana H. i Varela F. (1980). *Autopoiesis and Cognition – the Realization of Human Understanding*. Boston, MA: Shambhala.

Maturana H. i Varela F. (1992), *The Tree of Knowledge—The Biological Roots of Human Understanding*. Shambhala, Massachusetts.

McCarthy, J. i Hayes, P. J. (1969). Some Philosophical Problems from the Standpoint of Artificial Intelligence. W: B. Meltzer, D. Michie (red.), *Machine Intelligence* 4, 463–502.

Morse, A. F. i Ziemke, T. (2008). On the role(s) of modeling in cognitive science. *Pragmatics & Cognition*, 16 (1), 37–56.

Nida-Rümelin, M. (2018). Freedom and the Phenomenology of Agency. *Erkenn* 83:61–87.

Noë, A. (2004). *Action in perception*. Cambridge MA: MIT Press.

Nogueira, Y.L.B., de Brito, C.E.F., Vidal, C.A. i Canvalcante-Neto, J.B. (2019). Towards intrinsic autonomy through evolutionary computation. *Artificial Intelligence Review*, vol. 53, 4449–4473.

Moreno, A., Etxeberria, A. i Umerez, J. (2008). The autonomy of biological individuals and artificial models. *BioSystems* 91, 309–319.

O'Connor, T. (2009). Agent-Causal Power. W: T. Handfield (red.), *Dispositions and Causes*. Oxford: Oxford University Press, 184–214.

Olson, E.T. (2011). The Extended Self. *Minds & Machines* 21, 481–495.

Russell, S. (2019). *Human Compatible: AI and the Problem of Control*. UK: Allen Lane.

Sparrow, R. (2021). Why machines cannot be moral. *AI & Society,* 36, 685–693.

Taylor, Ch. (2001). Pojęcie osoby. W: J. Górnicka-Kalinowska (red.), *Filozofia podmiotu* (tłum. R. Wieczorek). Warszawa: Fundacja Aletheia.

Thompson, E. (2007). *Mind in life: biology, phenomenology and the sciences of mind*, 1st edn. Cambridge: Harvard University Press.

Tomczyk, B. (2022). *Podmiotowość rozszerzonych systemów poznawczych*. Lublin: Wydawnictwo UMCS.

Tomczyk, B. (2024). The Locus of Agency in Extended Cognitive Systems. *Journal for General Philosophy of Science*, 55. https://doi.org/10.1007/s10838-023-09666-1

Tomczyk, B. (2024). I Act Therefore I Live? Autopoiesis, Sensorimotor Autonomy, and Extended Agency. *Diametros*, May, 1–22. https://doi.org/10.33392/diam.1847

Woolford, F.M.G., Egbert, M.D. (2022). Goal oriented behavior with a habit-based adaptive sensorimotor map network. *Front Neurorobot.* 16. doi: 10.3389/fnbot.-2022.846693.

van de Poel, I. (2020). Embedding Values in Artificial Intelligence (AI) Systems. *Minds & Machines,* 30, 385–409.

Varela F. (1979), *Principles of Biological Autonomy*. New York: North Holland.

Varela F., Thompson E. i Rosch E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press.

Woolford, F.M.G. i Egbert, M.D. (2022). Goal Oriented Behavior With a Habit-Based Adaptive Sensorimotor Map Network. *Frontiers in Neurorobotics*, May 10; 16:846693. PMID: 35619969; PMCID: PMC9127740

Xie, Z. i Jin, Y. (2018). An Extended Reinforcement Learning Framework to Model Cognitive Development With Enactive Pattern Representation. W: *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 3, 738–750.

Reflective Agency vs. Sensorimotor Autonomy in the Debate on the Agency of Artificial Cognitive Systems

Abstract: In the presented article, I invoke two concepts of agency: the first, formulated within the framework of analytical philosophy of intentional action, which assumes the reflexivity of the agent, and the second, developed within the enactivist conception of agency as sensorimotor autonomy. I place these two concepts of autonomous action in the context of the discussion on the agency of artificial systems, which are currently gaining increasing independence in terms of selecting methods for solving specific tasks. In contrast to my earlier arguments expressed in the book "*The Agency of Extended Cognitive Systems"*, I highlight the explanatory advantage of the enactivist conception, which, although it cannot fulfill some of the functions of the standard definition od agency, proves more useful in discussions about the moral implications of artificial intelligence development. These implications are currently being actively discussed within the framework of the *Value Alignment Problem*, which concerns the alignment of human values with those that guide, and in particular will guide in the future, artificial intelligence. In this paper, I also point to the applications of the enactivist notion of agency in explaining the functioning of group systems as agents of action.

Keywords: autonomy, agency, agent, enactivism, artificial intelligence

Dr hab. **Barbara Tomczyk** jest adiunktką w Katedrze Logiki i Kognitywistyki Uniwersytetu Marii Curie-Skłodowskiej w Lublinie. Jej zainteresowania badawcze skupiają się na filozoficznych konsekwencjach stosowania różnego rodzaju artefaktów jako wzmocnień poznawczych. W szczególności, jej badania dotyczą problematyki sprawczości i odpowiedzialności w kontekście rozszerzonych, rozproszonych i sztucznych systemów poznawczych. Autorka szeregu publikacji naukowych w tym zakresie, a także wykonawczyni grantów Narodowego Programu Rozwoju Humanistyki MNiSW oraz "Sonata" NCN.