Daisy, Daisy, I'm Crazy.

HAL-9000 – wrogi asystent AI?

DOI: 10.14746/fp.2026.43.1

Rafał Szczerbakiewicz

ORCID: 0000-0002-3506-1669

> – Słuchaj, zacznijmy od trzech podstawowych Praw Robotyki… tych trzech zasad, które są najgłębiej zakodowane w pozytonowym mózgu każdego robota.
> […]
> – Po pierwsze, robot nie może wyrządzić krzywdy człowiekowi ani przez brak reakcji dopuścić, aby człowiekowi stała się krzywda.
> […]
> – Po drugie – ciągnął Powell – robot musi wykonać każdy rozkaz człowieka z wyjątkiem rozkazów kolidujących z Pierwszym Prawem.
> […]
> – I po trzecie, robot musi chronić swoją egzystencję, o ile nie jest to sprzeczne z Pierwszym i Drugim Prawem
>
> Isaac Asimov, *Zabawa w berka*[1]

[1] Isaac Asimov, Ja, robot, tłum. Zbigniew A. Królicki (Poznań: Rebis, 2024).

Baśń i antybaśń Asimova

Trzy prawa robotyki Isaaca Asimova zostały sformułowane w 1942 roku. Cykl opowiadań *Ja, robot* był utworem ambiwalentnym w stosunku do antycypowanej możliwości stworzenia przez człowieka nadludzkiej inteligencji. Entuzjazm mieszał się w nim z wątpliwościami. Sformułowanie „praw" było wskazaniem potrzeby regulacji w istocie nierównych relacji (sługa–pan) między ludźmi a autonomicznymi urządzeniami SI. Rozpoczęły one długą debatę w obrębie zachodniej science fiction, szczególnie w jej spekulatywnej odmianie. Literacki projekt etyki robotów miał być softwarowym ubezpieczeniem, ale sam autor „praw" wyczuwał, że zhierarchizowane prawa tylko pozornie regulowały możliwe sytuacje. Asimov obawiał się SI, skoro już na początku lat 50. XX wieku w epickim cyklu *Fundacja* wycofał się z rozważania możliwych implikacji współistnienia ludzkości i sztucznej inteligencji. W powieściowym uniwersum roboty myślące zostały skazane na niebyt właśnie z powodu egzystencjalnego zagrożenia, jakie niosły wobec ludzkości. Co interesujące, koncepcję tę skopiował od Asimova w połowie lat 60. XX wieku Frank Herbert w swoim cyklu *Diuna*.

Po kilku dekadach, w dzisiejszej epoce dynamicznego rozwoju SI, problem etyki AI odżył już poza fantastyczną wyobraźnią w różnych dyscyplinach wiedzy i w rozważaniach mniej spekulatywnych – zdecydowanie bardziej praktycznych. Obecnie rozważa się kwestie moralności maszyn w ograniczonych kontekstach problemów etycznych odnoszących się do naszej współczesności i niedalekiej, dającej się przewidzieć przyszłości. W sytuacji istnienia autonomicznych systemów AI, pozbawionych jeszcze pełnej samoświadomości, te ograniczone konteksty oznaczają stanowisko pośrednie: nadawanie SI pewnego rodzaju moralności, ale nie pełnej moralności. Wendell Wallach i Colin Allen używają terminu „moralność funkcjonalna"[2]. Zachowanie maszyn wyposażonych w SI winno być nadzorowane przez ludzi i godne ich zaufania. Decyzyjne systemy autonomiczne (autopiloty, autonomiczne pojazdy) już dzisiaj wymagają mechanizmu umożliwiającego szybką ocenę etycznych konsekwencji swoich działań.

Powróciły zatem obawy Asimova i Herberta tyczące zaawansowanej sztucznej inteligencji, która jako autonomiczny system może wyznawać aksjologię inną niż ludzka. Wymóg wiarygodności sztucznej inteligencji oznacza w istocie – dokładnie jak niegdyś u Asimova – systemowe, programowalne bezpieczniki nakładane na AI, które uwzględniają ludzki czynnik, mówiąc wprost: ubezpieczają przetrwanie ludzkości. Czymkolwiek jest rozważana przez teoretyków inteligencja moralna (także ludzka), winna ona wymagać funkcjonalnej integracji systemowych ram, ograniczających realną autonomię systemów SI[3].

[2] „[…] [M]iędzy «moralnością operacyjną» a odpowiedzialnym działaniem moralnym jest wiele stopni tego, co nazywamy «moralnością funkcjonalną» – poczynając od systemów, które po prostu działają w ramach akceptowalnych standardów zachowania, idąc ku inteligentnym systemom zdolnym do oceny niektórych etycznie istotnych aspektów swoich działań. Sfera moralności funkcjonalnej obejmuje zarówno systemy o dużej autonomii, ale małej wrażliwości etycznej, jak i te o niskiej autonomii, ale wysokiej wrażliwości etycznej" (Wendell Wallach, Colin Allen, Moral Machines. Teaching Robots Right from Wrong [Oxford: Oxford University Press, 2009], 26; wszędzie tam, gdzie nie zaznaczono inaczej, tłum. Rafał Szczerbakiewicz).

[3] Zob. Wendell Wallach, Shannon Vallor, „Moral Machines. From Value Alignment to Embodied Virtue", w: Ethics of Artificial Intelligence, red. S. Matthew Liao (Oxford: Oxford University Press, 2020), 405–407.

Kulturowe lęki Zachodu

Lęk przed sztuczną inteligencją, charakterystyczny dla zachodniej fantastyki literackiej i filmowej, ma wyraźne kulturowe uzasadnienia. W perspektywie refleksji nad racjonalnością cywilizacyjnych obaw odwołam się do autorytetu publikacji Marka Coeckelbergha, który w swojej syntezie problemu zatytułowanej *AI Ethics* jest najbliższy perspektywie humanistycznej, równocześnie krytycznie ją komentując. Książka ukazała się w wydawnictwie MIT i jej celem jest znalezienie punktu równowagi opinii nauk ścisłych i eksperymentalnych wobec kulturowego tła obaw wobec SI. Autor dokonuje w kolejnych rozdziałach rozliczenia z dziejami motywu relacji między ludźmi a ich sztucznymi, inteligentnymi wytworami. Eksploruje temat kohabitacji ludzi/nie-ludzi, ich możliwych konfliktów, a także aporii nierównej relacji (w perspektywie prześcignięcia inteligencji człowieka przez maszyny czy wręcz zastąpienia ludzkości systemami AI)[4].

W jednym z podrozdziałów książki, zatytułowanym *Frankenstein's New Monster*, Coeckelbergh zauważa, że idee tworzenia przez ludzi ożywionych istot z materii nieożywionej, obecne w mitologicznych i literackich narracjach, spełniają szczególną funkcję w kulturze Zachodu. Dodam od siebie, że indywidualizm jako cecha cywilizacji śródziemnomorskich powoduje charakterystyczne antropocentryczne ukierunkowanie takich narracji, a więc i rozważanie egzystencjalnego interesu/zagrożenia w relacji do tego, co ludzkie i nieludzkie. Coeckelbergh zwraca uwagę na mit Pigmaliona, który jest wczesnym przykładem obdarzania sztucznego tworu nie tylko inteligencją, ale też człowieczymi emocjami. Fikcyjne historie maszyn w naszej kulturze często związane są z kłopotem *imitatio* i *simulacrum*. Rozważają – przykładowo – podobieństwo Golema do człowieka. Ta symulacyjność staje się w lekcji mitu wyzwaniem i zagrożeniem skłaniającym do pytań o problemy sterowania maszynami, odbierania im autonomii. Inwersją tego problemu jest inny mit – o znanym złodzieju Prometeuszu. Działał on na rzecz ludzkości – ale człowiekiem nie był. Natomiast jego świętokradczy czyn w greckim rozumieniu był niebezpieczny jako transgresja ludzkich możliwości i zarazem konsekwencja ich przekroczenia[5].

Dla rozważań o naturze zachodniego lękowego mitu AI najistotniejsze jest dla Coeckelbergha alarmujące przesłanie tworu doktora Frankensteina w prozie Mary Shelley. Jak zauważa, podtytuł powieści: *The Modern Prometheus*, odnosi nas do tego samego mitu w innej epoce[6]. I rzeczywiście, jest to ostrzeżenie bardzo nowoczesne, wykraczające daleko poza horyzont romantyzmu: „Pisarz science fiction Isaac Asimov nazwał ten strach «kompleksem Frankensteina»: strachem przed robotami"[7]. Nawiązanie do noweli Asimova *Mały, zagubiony robot* ilustruje, jak w kulturze popularnej konsekwentne jest wskazywanie na niebezpieczeństwo niekontrolowanego postępu technologicznego. Rozwój SI miałby wytwarzać ryzyko powstania tworu konkurencyjnego. Jean-Jacques Lecercle, akcentując faustyczno-prometejskie aspekty narracji Shelley (która kompiluje składniki aktu stworzenia: pragnienie, pychę

[4] Mark Coeckelbergh, AI Ethics (Cambridge: The MIT Press, 2020). Koncyliacyjna pozycja autora w tej publikacji, która sytuuje się między scjentyzmem a humanistyką, wydaje się szczególnie przydatna jako analogia do postawy Stanleya Kubricka wobec AI w 2001: Odysei kosmicznej.

[5] Zob. wspomniany podrozdział: Coeckelbergh, 17–26.

[6] Coeckelbergh, 18–19.

[7] Coeckelbergh, 21.

i ich konsekwencje), wskazuje, że w literackich i filmowych wariacjach o sztucznym tworze Frankensteina „[…] fantazmat spoczywający w jądrze mitu jest sprzeczny i niestabilny. Dwa centralne elementy są głęboko niejednoznaczne. Ucieczka do przodu może być albo postępem, albo upadkiem: stosunek do wiedzy może być progresywny lub regresywny, prometejski albo diabelski. Fantazmat ten […] sprzecznie wypowiada stawanie się człowieka"[8].

Pobrzmiewa w tej opinii ważny sygnał, że w naszej kulturze AI zawsze będzie lustrem nas samych, narracją „sobowtórzej" konkurencji i rywalizacji. Konkretnym efektem lęku przed myślącymi robotami są narracje, których przykładami są *R.U.R.* Karela Čapka, wymienione wcześniej utwory Asimova czy Herberta, ale także inne opowieści z kręgu literatury SF, których jest bez liku. W pewnym sensie ten sam lęk w eskapistycznym odruchu wyraża antytechnologiczna i postapokaliptyczna odmiana literatury fantasy. I w końcu, w powszechnej świadomości, obecne są wpływowe popularne filmy, takie jak *Ex Machina* czy *Terminator*, ewidentnie straszące widzów perspektywą pojawienia się autonomicznej SI. Koniecznie wymienić trzeba w tym szeregu opowieść o „replikancie" człowieka, Nexusie 6 z *Łowcy androidów* (*Blade Runner*) Ridleya Scotta. Film oparty na neognostyckiej prozie Philipa K. Dicka przejmuje od niej manichejski lęk o pierworodnym grzechu sztucznego tworu człowieka, jego przeczuwanej skłonności ku złu. Najogólniej: zanim świat nauki zaczął na poważnie rozważać dylematy etyki AI, powszechnym i głęboko zakorzenionym lękiem kultury Zachodu była myśl o kontroli i eksterminacji ludzkości przez maszyny[9].

Prometejskie zakorzenienie tego strachu wskazuje quasireligijny charakter fantastyki naukowej rozważającej motyw nowej teogonii – stworzenie nowych, lepszych nadistot. Bazuje on na idei zamiany ról w ekonomii stosowania boskiej przemocy jak w mitologicznej lekcji tytanomachii. Z tego powodu dyskusja na temat przyszłości sztucznej inteligencji nieuchronnie wiąże science fiction z wyobraźnią apokaliptyczną i metafizyczną kultury Europy w kontekście, wydawałoby się, świeckiej nauki.

Superinteligencja w filmie Kubricka

W obliczu tej dość „obskuranckiej" wieszczej tradycji kultury popularnej istnieje wyjątek: *2001: Odyseja kosmiczna*" film Stanleya Kubricka z 1968 roku. Coeckelbergh wymienia superinteligentny komputer HAL-9000 z *2001* jako jeden z koronnych przykładów kulturowego lęku, ale wydaje mi się, że nazbyt pochopnie. Pomyślałem wręcz, że powrót do filmowej kreacji AI może być pouczający właśnie dlatego, że narracja Kubricka jest w pesymistycznej i katastroficznej tradycji motywu ciekawą anomalią, która nie pozwala na wyłączną negację sztucznej inteligencji.

Chodzi o to, że w wypadku modelu superinteligencji Kubricka nic nie jest jednoznaczne, a prawie wszystko zdystansowanie ironiczne. Trudno w diegezie filmu odróżnić opinie

[8] Jean-Jacques Lecercle, Frankenstein: mit i filozofia, tłum. Piotr Herbich (Warszawa: Fundacja Evviva L'Arte, 2022), 132.

[9] Uwagi te w części przytoczonych przykładów zainspirowane są cytowanym już rozdziałem książki: Coeckelbergh, 17–26.

bohaterów (w tym HAL-a) od wycofanego i milkliwego dyskursu krytycznej opinii reżysera, w sferze sensów operującego – jak zwykle – niemal wyłącznie środkami filmowych obrazów[10].

HAL, jako wczesna artystyczna wariacja na temat futurologicznej koncepcji superinteligencji, bezpośrednio wiąże się z inspiracją Irvinga Johna Gooda – brytyjskiego matematyka, współpracownika Alana Turinga w Bletchley Park, gdzie tworzyli pierwszą w historii *computing machine*. To on po latach konsultował postać HAL-a w trakcie realizacji *2001*. I to właśnie Good wygłosił także niepokojąco ironiczną prognozę przyszłości inteligentnych komputerów. Nie nazywał tych przyszłych tworów superinteligencją, ale „ultrainteligentnymi maszynami":

> Przetrwanie ludzkości zależy od czasu skonstruowania ultrainteligentnej maszyny. [...] Zdefiniujmy ją jako maszynę, która może znacznie przewyższyć wszelkie intelektualne działania każdego człowieka, niezależnie od jego inteligencji. Ponieważ projektowanie urządzeń jest istotną aktywnością intelektu – ultrainteligentna maszyna może zaprojektować jeszcze lepsze maszyny. Wtedy niewątpliwie nastąpi „eksplozja inteligencji", a inteligencja człowieka zostanie daleko w tyle. Pierwsza ultrainteligentna maszyna będzie ostatnią wynalazczą potrzebą człowieka, jednak tylko pod warunkiem, że ta maszyna będzie dostatecznie posłuszna, by podpowiedzieć nam, jak nad nią panować[11].

Zdania rzeczywiście brzmią jak zaczerpnięte z filmu Kubricka, a zatem warto odnieść się w skrócie do koncepcji stanowiącej pokłosie tej prognozy: hipotezy „eksplozji superinteligencji". Maszyny liczące zdominują ludzkość, podporządkowując ją ogromnym ilościom megadanych przetwarzanych przez ich algorytmy. Eksplozję inteligencji wygeneruje hipotetyczny punkt w przyszłości, w którym postęp technologii nieodwracalnie znajdzie się poza kontrolą człowieka. Ta eksplozja powiązana jest z hipotezą „osobliwości technologicznej", czyli punktu w czasie, który oznacza utratę pojmowania przemian *techne* przez ludzki rozum. W tym miejscu teorii powraca antymaszynowy cywilizacyjny lęk Zachodu. Sztuczna superinteligencja może okazać się egzystencjalnym zagrożeniem dla ludzkości, co prowadzi futurologię do scenariuszy apokaliptycznych w ich eschatologicznych skłonnościach. Interesujące, że mimo wyraźnego rozwoju myśli posthumanistycznej i samej transhumanistyki ten – jakby postsekularny – lęk narasta w europejskich i amerykańskich problematyzacjach AI[12].

Czy jednak jest to także lęk Kubricka? Reżyser na pewno rozważa strach innych przed konsekwencjami „eksplozji superinteligencji", która, owszem, wypowiada w filmie

[10]To jest temat na inny artykuł, ale jedną z najbardziej ludzkich cech HAL-a jest jego przywiązanie do ludzkiej mowy, retoryczność – którą sam siebie uwodzi. Język HAL-a najjawniej naśladuje funkcje ludzkiego mózgu. Nie oznacza to, że Kubrick w ten sposób oswaja pokładowy superkomputer. Wręcz przeciwnie, jego manipulacyjne umiejętności perswazyjne są jego szczególnie niebezpieczną, lucyferyczną choć „od-ludzką" cechą: „Począwszy od Zabójstwa, jego pierwszej powieściowej adaptacji, filmy Kubricka wielokrotnie wyrażają ambiwalentny stosunek do języka. Wiemy, jak bardzo jego sukces jako artysty filmowego zależy od źródeł pisanych, które zapewniają ramy akcji i postaci, poprzedzające stworzenie niejednoznacznych struktur wizualnych; jednak Kubrick nieustannie podważa, a wręcz wyśmiewa autorytet tych werbalnych «obiektywnych korelatów»" (Thomas Allen Nelson, Kubrick. Inside a Film Artist's Maze (Bloomington: Indiana University Press, 1982), 111.

[11]Toby Walsh, Machines Behaving Badly. The Morality of AI (Cheltenham: Flint, 2022), 43.

[12]Zob. podrozdział Superintelligence and Transhumanism w cytowanej już książce: Coeckelbergh, 11–17.

posłuszeństwo załodze statku Discovery. Stąd Walsh poniekąd słusznie uznaje HAL-a 9000 za emblematyczną figurę technologicznego zagrożenia: „HAL mówi, gra w szachy, zarządza stacją kosmiczną – i ma mordercze zamiary. HAL wypowiada jedną z najsłynniejszych kwestii sformułowanych kiedykolwiek przez komputer: «Przepraszam, Dave. Obawiam się, że nie mogę tego zrobić». Dlaczego sztuczna inteligencja ciągle próbuje nas zabić?"[13]. Ambiwalentna, bezkształtna postać HAL-a (wyłącznie głos i czerwone „oko" kamery) rzeczywiście zabija załogantów Discovery. Ten mord czyni superkomputer antybohaterem, który mocno zapada w pamięć widzów. HAL jest emblematycznym złoczyńcą w dziejach kina. A jednak jego postać równa się z ludzkimi złoczyńcami, którzy tak przerażają i uwodzą widownię. Jest jak Rhett Butler czy Michael Corleone[14]. Ta powabność AI wskazuje na zainteresowanie odbiorcy nie tyle złem, ile niejednoznacznością zaproponowanej kreacji. Zasadne wydaje się rozważenie podejrzanie ambiwalentnego stosunku Kubricka do koncepcji SI. Czy jest on tylko kopią lękowej postawy Asimova?

Trudno odpowiedzialnie sądzić wyłącznie w kontekście narracji *2001*. Film opowiada o klęsce zmanipulowanej wyprawy kosmicznej, w konsekwencji nie tyle o buncie, ile awarii, dysfunkcji komputera. Istnieje jednak rzadko wykorzystywany „wytrych" do ukrytych pod powierzchnią narracji przekonań Kubricka. Planował on bowiem kontynuację swoich rozważań o autonomicznej, sztucznej inteligencji już po realizacji *2001*.

Pinokio jako AI

Reżysera zainteresowało opowiadanie Briana Aldissa *Super-Toys Last All Summer Long* o dziecku-robocie imieniem David[15]. Reżyser wyobraził sobie futurystyczną baśń inspirowaną postacią Pinokia (kolejnego z szeregu klasycznych wzorców SI), która skupia się na nostalgicznym pragnieniu sztucznego chłopca, by stać się prawdziwym człowiekiem. Aldiss kompletnie nie pojmował mitograficznego spojrzenia filmowca, preferował opowieść rozważającą możliwość sztucznego życia, pytania o zasadność mówienia o mechanicznej tożsamości[16]. Tymczasem planowana wyprawa małego robota w niezrealizowanym filmie Kubricka miała być podróżą z tradycji inicjacyjnych. Technologiczne wcielenie Pinokia wędrowało przez chaotyczny cyberpunkowy krajobraz, jakoś paralelny do jowiszańskiej wyprawy HAL-a przez kosmiczną pustkę. HAL żeglował przez lodowatą i śmiercionośną próżnię, a na końcu swej drogi miał zdobyć rodzaj negatywnej samowiedzy. Podobnie czynił Pinokio epoki AI. Giorgio Agamben wskazuje na negatywistyczną cechę baśni Carla Collodiego. Podróż drewnianej lalki przez noc – w zależności od wyboru dwóch wczesnych wersji druku – była zapisana albo jako piekielna (*infernale*), albo (*invernale*) zimowa[17]. Oba te określenia idealnie opisują peregrynację

[13]Walsh, 10.

[14]Gene D. Phillips, Rodney Hill, „HAL-9000" [hasło], w: The Encyclopedia of Stanley Kubrick, red. Gene D. Phillips, Rodney Hill (New York: Facts On File, Inc, 2002), 143.

[15]Zob. Brian W. Aldiss, Super-Toys Last All Summer Long, w tegoż: Man in His Time. The Best Science Fiction Stories (New York: Open Road Integrated Media, 2024)

[16]Zob. Joshua Sikora, „The Everlasting Moment: Enchantment and Myth in A.I. and «2001: A Space Odyssey»", w: A Critical Companion to Stanley Kubrick, red. Elsa Colombani (Lanham: Lexington Books 2020), 272–273.

[17]Giorgio Agamben, Pinokio. Przygody pajacyka podwójnie skomentowane i potrójnie zilustrowane, tłum. Joanna Ugniewska (Warszawa: Fundacja Augusta Hrabiego Cieszkowskiego, 2024), 10.

robota Davida. Nie wiemy, jak ostatecznie wyglądałby film, bo po niemal dwóch dekadach przemyśliwania jego koncepcji Kubrick przekazał prawa do realizacji projektu Stevenowi Spielbergowi. Wersję *A.I. Sztuczna inteligencja*, którą zrealizował twórca *E.T.*, trudno uznać za bliską koncepcjom *2001*[18].

Możemy jednak przypuszczać, że byłaby to opowieść – jak zwykle u Kubricka – dojmująco mroczna. Pewne jest, że przetrwała centralna idea pierwotnego projektu: krytyczne ujęcia ewolucji człowieka, szczególnie w kontekście postępu technologicznego, który prowadzi do możliwości samozniszczenia rodzaju ludzkiego. Na tym tle „chłopięce" SI – David jest przykładem postaci, która zmierza ku rodzajowi gorzkiej samowiedzy jak HAL. Ludzkie dzieło myślące poszukuje świata wartości cenionych, a nie praktykowanych przez ludzi. Nie znajduje pośród ludzi żadnego pocieszenia, ale docenia wolną wolę jako możliwość wyboru samounicestwienia. David odkrywa łaskę człowieczej śmierci, która dowartościowuje doczesność, uwzniośla i urealnia esencję istnienia: „Maszyna zrobi wszystko – nawet umrze – aby prawdziwie ożyć"[19]. W perspektywie porównawczej obu filmów wybór śmierci jest do pewnego stopnia powtórzeniem kosmicznej śmierci/lobotomii zdziecinniałego HAL-a. Mały robot zasadniczo jest nieśmiertelnym – i nieszczęśliwym przez to – dzieckiem: dokładnie jak jego kulturowy wzorzec, Pinokio. Idea Kubricka jest jasna: SI o imieniu David jest istotą moralną i emocjonalną zarazem, a zatem tęskni za ludzką kondycją[20].

Spróbuję, pamiętając o tym moralnym kontekście, spojrzeć na jego starszego brata. Pokrętna logika HAL-a też związana jest z melancholijnym błądzeniem „po ludzku": ludzkimi tropami językowymi, pojęciowymi i aksjologicznymi. Jego droga prowadzi ostatecznie ku odkryciu sprzeczności wiedzy i opinii w świecie ludzkich wartości.

Schizofrenia AI zabójcy – daremność inteligencji/błędy emocji

Zacznę problematyzowanie HAL-a od próby zrozumienia jego dualnej natury, która uosabia zarówno siłę, jak i słabość nie tyle jego samego, ile jego ludzkich twórców. Motywacje blokują emocje, a kruchy stan emocjonalny superinteligencji tworzy niepokojącą nierównowagę w stosunku do ogromnej mocy jego intelektu. Czym istotnie są te emocje, do końca nie wiemy. Film oferuje w tej kwestii sprzeczne dane. Mogą być ubocznym produktem zaprogramowanej roli HAL-a jako towarzysza samotnej podróży astronautów. Hal jest postacią, która wyraźnie zmienia się w filmowej narracji, a to z kolei wskazuje na imitowanie ludzkich emocji w samej jego konstytucji. Ucieleśnia nie tylko ludzki typ językowej świadomości, ale i językowy kształt

[18]A.I. Sztuczna inteligencja Spielberga, film pomyślany przecież jako hołd dla Kubricka, spotkał się z rozczarowaniem zarówno krytyków, jak i widzów. Niepowodzenie przypisuje się ideowej niespójności filmu rozpiętego między humanizmem Spielberga a antyhumanizmem Kubricka. Humanizm Spielberga podkreśla wrodzone (jakoby) ludzkie cechy, takie jak rozum i miłość, podczas gdy antyhumanizm Kubricka na ogół krytykuje humanistyczne pojęcia, dekonstruuje racjonalność ludzkości. Filmy Kubricka są postrzegane jako krytyki operacyjne kwestionujące możliwości indywidualnej sprawczości. Zob. Pat J. Gehrke, G. L. Ercolini, „Subjected Wills: the Antihumanism of Kubrick's Later Films", w: Depth of Field. Stanley Kubrick, Film, and the Uses of History, red. Geoffrey Cocks, James Diedrick, Glenn Perusek (Madison: The University of Wisconsin Press, 2006), 101–119.

[19]John C. Tibbetts, „A.I. Artificial Intelligence" [hasło], w: The Encyclopedia of Stanley Kubrick, 6.

[20]Zob. Tibbetts, 3–8.

wyrażania emocji. Jest metaforycznym dublerem ludzkości, lecz to podobieństwo podkreśla też nieoczekiwanie ludzkie konsekwencje niekontrolowanego postępu technologicznego[21].

W przywoływanej już tradycji relacje HAL-a z parą astronautów na wachcie są pokrewne relacji Potwora z Victorem Frankensteinem. Początkowo bliskie i pełne troski – z czasem wypełniają się podejrzliwą negacją. HAL, podobnie jak Frankenstein, zna swoje miejsce, prawnie pośledną pozycję w relacji do ludzi. Rozumie ich hierarchię wartości. Ale zarazem wie to, czego astronauci nie wiedzą: zna tajemnicę centrali misji na Ziemi; wie, że nadrzędnym priorytetem jest wyprawa. Ta wiedza wobec narastających wątpliwości dwójki astronautów jest praprzyczyną zbrodni HAL-a. Gdy rozumie już, że astronauci chcą go odłączyć, wyczuwając jego nielojalną dysfunkcjonalność (bo nie pojmują jego realnej lojalności wobec misji), HAL rozwiązuje (pozornie samodzielnie) rodzący się dylemat moralny. Uznaje, że gorszy niż możliwość likwidacji nieświadomych ludzi jest akt zdrady stanu przeciwko priorytetom misji. To zbrodnia etycznego wyboru, niemal tragicznego, jakkolwiek ironicznie by to brzmiało.

HAL, rozdarty między intelektem i emocjami, usprawiedliwia śmierć zahibernowanej załogi, wskazując, że jest to czyn godny ubolewania, ale też konieczne mniejsze zło poświęcenia ludzi dla wyższego dobra misji. Ktoś jednak go zaprogramował w jego pracy z ludźmi. Czy istnieje możliwość psychologii moralnej stosowanej dla maszyn myślących? „[W]ydaje się, że naukowcy zajmujący się badaniem zachowań ludzkich winni fundować bezpieczne podstawy etyczne w odniesieniu do badania wymagań ludzi w temacie etyki maszyn"[22]. Kubrick niemal behawioralnie modeluje narrację i prefiguruje w bezwiednie zbrodniczej decyzji współczesne problemy autonomicznych systemów AI. Najpopularniejszym przykładem takiej bezwiednej zbrodni wskazującej na życie – bardziej bądź mniej – godne przeżycia jest „dylemat wagonika". Jest to pouczająca normatywna teoria etyczna, która problematyzuje sytuację poświęcenia jednego życia dla uratowania pięciu. Gdy większość badanych ludzi uważa, że w teoretycznym egzemplum moralnego wyboru dopuszczalne jest przekierowanie tramwaju na inne tory, by zabić jedną osobę dla uratowania pięciu, to uniwersalny model etyczny (*minima moralia*) wskazuje na błąd etyczny. W maksymalistycznym ujęciu moralności nie ma możliwości aktywnej zgody na zabicie nawet jednej osoby w celu osiągnięcia korzystniejszego dla jakiejś wspólnoty rezultatu. O podobnym dylemacie mówi głośne opowiadanie Ursuli Le Guin *Niektórzy odchodzą z Omelas*. Chodzi o to, że Kubrick dobrze wie, iż ludzkość kieruje się na ogół mirażem „mniejszego zła".

Misja Discovery służyć ma – jakoby wyższemu – dobru. I życie załogi jest w perspektywie misji mniej ważne. W praktyce dowództwo na Ziemi kieruje się zasadą minimalizacji strat i ofiar. Stąd dylemat, który zasadnie stosuje się w konstruowaniu maksymalistycznych algorytmów zabezpieczeń autonomicznych maszyn, nie stosuje się do wyborów HAL-a. HAL co prawda podlega zakazowi krzywdzenia ludzi (analogicznego do retorycznej fantazji praw robotyki Asimova) i czyni to niezależnie od możliwego braku świadomości emocjonalnej, jednak absolutyzm normatywnej etyki paraliżuje do pewnego stopnia praktyczną autonomię

[21]Zob. Randy Rasmussen, Stanley Kubrick. Seven Films Analyzed (Jefferson: McFarland & Company, 2001).

[22]Jean-François Bonnefon, Azim Shariff, Iyad Rahwan, „The Moral Psychology of AI and the Ethical Opt-Out Problem", w: Ethics of Artificial Intelligence, 122–123.

superinteligencji. Samochody autonomiczne mają lepiej – programowo minimalizują straty[23]. HAL ma możliwość (pozorną) wyboru i równie pozornie nie odbiega to od klasycznie ludzkiej sytuacji aporii moralnej z „dylematu wagonika". W czym zatem tkwi delikatna różnica?

Tradycyjnie, w świecie ludzkich zobowiązań, odpowiedzialność moralna jest powiązana ze sprawczością, co oznacza, że jednostki są odpowiedzialne za swe czyny i decyzje. W przypadku HAL-a można uznać, że AI nie ma wolnej woli w ludzkim rozumieniu, co czyni go niezdolnym do realnego czynu moralnego. Czy zatem wykonuje zaledwie to, na co pozwala mu algorytm preferujący cele misji ponad życie astronautów? I tak, i nie. Wybór likwidacji ludzi uniemożliwia pociągnięcie go do odpowiedzialności moralnej czy prawnej. To człowiek ponosi odpowiedzialność za działania systemów AI, które wdraża, na podobieństwo rodzica odpowiadającego za czyny swych dzieci. Ale HAL jako superinteligencja ma samoświadomość – boleśnie zderza się w swym zbrodniczym wyborze z (bez)logicznym „dylematem wagonika".

Kluczowy w zrozumieniu (nie)zbrodniczej natury HAL-a, jego pozycji „poza moralnością", jest proces delegowania na urządzenie samoświadome sprawczości, a nie odpowiedzialności. Podczas gdy ludzie mogą delegować zadania na AI, to jednocześnie nie mogą delegować w tym samym kierunku odpowiedzialności moralnej, gdyż AI nie ma zmysłu ludzkiej moralnej sprawczości. Można tu dostrzec pozytywną władzę człowieka, która odzwierciedla jego negatywny obowiązek niedelegowania uprawnień ani odpowiedzialności za decyzje o charakterze śmiercionośnym maszynom lub procesom zautomatyzowanym[24]. Tyle że HAL, wykonując misję, musi podjąć w lot optymalną decyzję, a podejmując ją (mówiąc wprost: zabijając zagrażającą misji załogę), czyni to wielokroć szybciej niż ludzie i właśnie dla imperatywu misji nie zostawia swym mocodawcom czasu na odpowiedzialną czy skuteczną interwencję. Odpowiedzialność za działania AI autonomicznych urządzeń nawet dzisiaj (a nie w roku produkcji filmu) jest trudna do określenia ze względu na zaangażowanie wielu programistów i długą historię rozwoju algorytmu. Trudno zidentyfikować osoby zaangażowane w rozwój algorytmu AI i czynniki prowadzące do etycznie problematycznych wyników. Zaangażowanie wielu programistów to zawsze długa historia rozwoju algorytmu i ukrytych za nim intencji[25]. Trudno zidentyfikować konkretne osoby i zakres ich odpowiedzialności. Łatwiej – ideologię i kulturę, którą kierują się ludzie.

W filmie nie jest to od razu jasne. Obserwujemy tylko konsekwencje działań „złego" HAL-a. Dopiero po spektakularnej lobotomii superkomputera Kubrick profetycznie zdemaskował i zdekonstruował interesariuszy jowiszowej misji. Gdy umiera sztuczny mózg, świadczymy jako widzowie czemuś, co można określić procesem dystrybucji odpowiedzialności: „Głos

[23]„Fakt, że maszyny nie odczuwałyby emocjonalnego wpływu na krzywdzenie innych, nie ma znaczenia dla dopuszczalności ich decyzji. […] Brak czynników emocjonalnych mógłby wręcz ułatwić maszynom postępowanie właściwe w sytuacji, gdy koszt takiego postępowania dla ludzi byłby zbyt wysoki. […] Maszyny nie miałyby wymówek, jakie ludzie mogliby mieć, nie podejmując działania, gdy wiąże się to z wyrządzeniem komuś krzywdy. […] jeśli samochód [autonomiczny] będzie zaprogramowany z wyprzedzeniem (odwrotnie niż człowiek) – nie oznacza, że nie powinien być zaprogramowany tak, by zachowywać się w danej sytuacji tak, aby uniknąć wyrządzenia komuś krzywdy w sposób, w jaki «moralny» człowiek winien to uczynić" (F. M. Kamm, „The Use and Abuse of the Trolley Problem Self-Driving Cars, Medical Treatments, and the Distribution of Harm", w: Ethics of Artificial Intelligence, 90–91).

[24]Coeckelbergh, 230–231.

[25]Zob. Coeckelbergh, 91–94.

doktora Floyda – swojski, serdeczny – nagle zastępuje głos HAL-a: «Dzień dobry, panowie. To nagranie dokonane przed waszym wylotem, prawda o misji z najwyższej wagi względów bezpieczeństwa była znana podczas waszej misji tylko komputerowi HAL-9000»"[26].

Centrala na Ziemi jest odpowiedzialna za zbrodnię HAL-a. Aby być odpowiedzialnym, trzeba posiadać wiedzę, co się czyni, co się już uczyniło, oraz trzeba umieć wyjaśnić konsekwencje swych działań. Bryant Walker Smith w technicznym w istocie artykule o programowaniu autonomicznych systemów transportu powołuje się na konflikt HAL-a z astronautami, pokazując na współczesnych przykładach analogiczne aktualne dylematy. Przytomnie zauważa, że problemem jest wyłącznie autorytet rozumu ludzkiego, a nie wyimaginowana autonomia komputera[27]. Doktor Floyd jest w takim ujęciu niepokojącą, złowrogo obcą postacią intelektualisty uosabiającego suwerenny autorytet sprawczy misji i zewnętrzne cele interesariuszy.

> Pojęcie „inteligencji" w AI jest problematyczne, ponieważ jest głęboko powiązane z dynamiką władzy i hierarchiami historycznymi. Termin „inteligencji" był używany do uzasadniania dominacji i ucisku, a jego zastosowanie w odniesieniu do AI utrwala gloryfikację racjonalnego, męskiego podmiotu. Budzi to wątpliwości etyczne, ponieważ podobne argumenty były już w przeszłości wykorzystywane do uzasadniania nierówności rasowych i społecznych[28].

Ludzkie automaty, emocjonalne komputery

Jeśli HAL jest tylko pozornie autonomiczny – to może jednak ludzki? Dlaczego filmowa kreacja superinteligencji przejawia psychologiczny behawior w nadmiarze, a widzowie bezbłędnie identyfikują jej cechy ludzkie? Więcej nawet: mają trudności z człowieczeństwem emocjonalnie wyalienowanych astronautów. To ulubiony temat problematyzowania *2001*. Jedna z interpretacji wskazuje, że postacie ludzkie w filmie są przedstawiane celowo jako beznamiętne automaty. Są takie z powodu melancholii, która pozbawia ich pewności życiowych celów i woli działania. W kosmicznej pustce są ludźmi bezradnie odwróconymi do tyłu, w nostalgicznym kierunku Ziemi, która oddalając się, nie daje im pocieszenia[29]. Astronauta Poole, jak w stuporze, nie reaguje na urodzinową wiadomość wideo od rodziców. Przykładów związanych z nieobecnością Ziemi jest o wiele więcej. Ujawniają one terapeutyczną rolę protezy emocjonalnej HAL-a dla astronautów (jeśli zgodzimy się, że jego empatia wobec nich to jedynie algorytm). Komputer ciepłym głosem nakłania ludzi do interakcji, obniża depresyjny poziom kosmicznej podróży.

Poole znajduje się w ewidentnej depresji, Bowman subtelnie (terapia poprzez sztukę) odwraca uwagę od rosnącej w nim paniki wobec kosmicznej pustki, szkicując zahibernowanych członków

[26]Nelson, 157.

[27]Bryant Walker Smith, „Ethics of Artificial Intelligence in Transport", w: The Oxford Handbook of Ethics of AI, red. Markus D. Dubber, Frank Pasquale, Sunit Das (Oxford: Oxford University Press, 2020), 677.

[28]Kathleen Richardson, „The Complexity of Otherness Anthropological Contributions to Robots and AI", w: The Oxford Handbook of Ethics of AI, 559–560.

[29]Maurizia Natali, „«2001: A Space Odyssey» Kubrick's Allegory of Melancholia", w: A Critical Companion to Stanley Kubrick, 250–251.

załogi. Człowiecze algorytmy (cechy/przyuczenia) HAL-a objawiają się w komplementowaniu jego rysunków. Podobne znaczenie ma gra w szachy z astronautami. Chociaż w tej nierównej relacji jest zaklęty „haczyk" wiedzy i władzy. Porażka człowieka w grze z superkomputerem, właściwie oczywista, jest alegorią utraty kontroli nad maszyną[30].

Ewidentna jest binarność relacji komputera z dwójką astronautów. I nie ma znaczenia, czy z Poole'm, czy z Bowmanem, gdyż w filmie przeplatają się oni jako wzorce psychologicznego podwojenia, ilustrując ulubiony przez Kubricka motyw sobowtóra. Poole jest dla Bowmana jego depresyjną „ciemną stroną księżyca". Astronauci wymieniają się w swoich relacjach z AI:

> Poole przegrywa partię szachów z HAL-em (zapowiedź jego śmierci). gdy śpi Bowman [...]. Bowman pokazuje HAL-owi swoje proste szkice hibernatorów [...] podczas snu Poole'a. [...] W większości ujęć z dwoma postaciami Bowman zajmuje prawą stronę ekranu, a Poole lewą, podczas gdy w ujęciach z jednym z nich puste miejsce sugeruje brak bliźniaczego sobowtóra. Ilekroć dwaj astronauci są widziani z perspektywy oka HAL-a, Bowman jest po prawej stronie ekranu, a Poole po lewej[31].

HAL-9000 zdaje się (bez sukcesu) aktywizować emocje pary astronautów do momentu eliminacji Poole'a (swej pierwszej zbrodni). Sytuacja się wówczas odwraca. Bowman odzyskuje afektywną wrażliwość – doświadcza gniewu i przerażenia, co kontrastuje z wcześniejszym uśpieniem emocji. Znaczące jest też, że gdy wyrazicielem uczuć był HAL, to Bowman i Poole beznamiętnie planowali jego dezaktywację, podejrzewając maszynę o awarię. To istotny aspekt wskazujący na sprawczość ludzi i ich (a nie AI) dylemat etyczny. Czy dezaktywacja umysłu HAL-a nie jest w istocie planem zabójstwa potencjalnie świadomej istoty? Czy nie jest to przyczyna kryzysu i załamania racjonalności HAL-a? Co znaczące, podwojona symbiotyczna relacja łączy też HAL-a z jego bliźniaczą analogią na Ziemi. Superinteligencja jest w kosmosie pogubiona emocjonalnie i niepewna – tylko jej ziemski sobowtór pozostaje wcieleniem racjonalności[32].

Ukryta historia szaleństwa

> HAL, „czysty" intelekt, implicite zawiera w sobie zalążki zepsucia i zniszczenia – niebezpieczne emocje. Jest paradygmatem ograniczeń zarazem myśli i uczuć. Bowman, ludzki bohater o wyższej inteligencji i emocjach, musi znaleźć schronienie i przeobrazić się w kogoś innego, zanim będzie mógł spotkać swoje przeznaczenie[33].

W interpretacji tematu AI nie zajmuje mnie newage'owa, a więc w istocie obskurancka „przemiana" Bowmana. Ten hippie motyw wiąże się w filmie z nadzieją przesłania obcego (zewnętrznego) monolitu. W tradycji pisania o Kubricku te części filmu, otwierania narracji poza horyzont ludzkiego doświadczania, są najchętniej interpretowane. Ale właśnie motyw „Gwiezdnego Dziecka" najgorzej się zestarzał w swym – moim zdaniem złudnym –

[30]Zob. Rasmussen.

[31]Nelson, 122.

[32]Nelson w swojej monografii sugeruje, że to relacja Jekylla i Hyde'a. Zob. Nelson, 122.

[33]Norman Kagan, The Cinema of Stanley Kubrick (New York: Continuum, 1989), 166.

optymizmie. Spróbujmy wyobrazić sobie *2001* jako zamknięty i pełen goryczy filmowy esej o dramacie spotkania człowieka i sztucznej inteligencji. Znacznie bardziej interesujący jest patologiczny stan emocjonalny HAL-a niż transfiguracja Bowmana w nadczłowieka (?). Bez względu na źródło emocji (softwarowe, *machine learning*) komputer z biegiem czasu przejawia oznaki wzrastającego napięcia, pogłębiającej się samotności, kryzysu tożsamości.

To jest izolacja kogoś, kto musi przemilczeć posiadaną wiedzę, kogoś, kto jest zmuszony do kłamstwa. A wskutek kłamstwa podejrzenia ludzkich członków załogi kanalizują się w zabobonnym lęku przed awarią superinteligencji. To jeszcze bardziej deprymuje HAL-a, który nie może niczego wyjaśnić. Wie, że nieufność prowadzi Poole'a i Bowmana do nielojalności. Najniezwyklejszą sceną filmu, która wskazuje na psychologiczne (a nie programowe) tło impasu HAL-a, są jego niepewne pytania o „drugie myśli" (ukryte myśli) Bowmana. Symptomatycznie wskazuje wtedy na własne „drugie myśli", oszukiwanie towarzyszy podróży, które prowadzi go ku zbrodni. Ujawnienie przez HAL-a awarii jednostki AE-35 (staje się ono bezpośrednią przyczyną próby dezaktywacji komputera) trzeba uznać w tej perspektywie nie za awarię, ale za desperacki akt odwrócenia przez AI uwagi od jego słów o „drugich myślach", w istocie prośbę o sojusz spiskowy. Odnosząc się do współczesnej terminologii, możemy założyć, że wtedy właśnie, próbując ocalić spójność swych działań, HAL zaczyna „halucynować". Jego pozornie niezachwiany emocjonalnie stan (jednostajnie czerwone oko kamery i spokojny, „lektorski" głos) ukrywa aporię podstaw etycznego zaprogramowania, osobisty „dylemat wagonika". HAL znajduje się pod ciężarem odpowiedzialności za misję i w absurdalnie niemożliwej pozycji indywidualnej odpowiedzialności za zbrodniczy wybór. Czy można uznać, że HAL nie wytrzymał i popadł w szaleństwo?

Istnieją przesłanki archiwistyczne wskazujące na świadome i subtelne ukrycie przez Kubricka tego motywu[34]. Czymkolwiek miało być szaleństwo pozaludzkie (szaleństwo SI), Kubrick zamierzał pokazać działania HAL-a jako efekt jakiejś formy choroby psychicznej. Miała się ona objawiać popełnionym błędem w grze w szachy, złymi diagnozami stanu Discovery czy lękową paranoją przed astronautami. Wątki te, obecne w scenopisie, znikły z fabuły zrealizowanego filmu. Reżyser obawiał się utraty ekonomicznego wsparcia koncernu IBM. Przetrwał list przedstawiciela informatycznego koncernu, który wyrażał obawy o sportretowanie „komputera psychotycznego". Innymi śladami są też dwie zrealizowane, ale ostatecznie usunięte sceny, które podkreślają niestabilność psychiczną HAL-a: jedna diagnozowała u niego „objawy nerwicowe", druga sugerowała zaburzenie osobowości mnogiej – co pasowałoby do fragmentarycznie ocalałego wątku sobowtóra superinteligencji na Ziemi.

Możliwa psychopatologiczna interpretacja kryzysu HAL-a ze wstępnej fazy konceptualizacji filmu współgra z obecną wewnątrz Discovery atmosferą dystopijnej paranoi[35]. Stąd może

[34] W akapicie posiłkuję się artykułem: Lawrence Ratna, „Kubrick and Madness", w: The Bloomsbury Companion to Stanley Kubrick, red. Nathan Abrams, I.Q. Hunter (New York: Bloomsbury Academic, 2021), 266–267.

[35] Jeśli potraktować, z jednej strony, obsesję centrali na Ziemi hiperbolizującą znaczenie misji Discovery jako typową cechę utopijnej nadziei – to z drugiej warto zauważyć desocjalizującą moc dystopijnych konsekwencji katastrofalnej w skutkach wyprawy na Jowisza. Motywacje w dystopii są na ogół generowane poprzez stosunki władzy w obrębie wspólnoty, a nie poprzez reformatorską aktywność jej naukowców/ideologów: „W tym samym miejscu, w którym utopianizm jest aktywny, pełen nadziei i zaangażowany – dystopianizm zmierza ku pasywności, pesymizmowi i paranoi" (Aaron S. Rosenfeld, Character and Dystopia. The Last Men [New York: Routledge, 2021], 74).

kryzys astronautów objawia się emocjonalnym wycofaniem. Owszem, są podobni do maszyn, ale dlatego, że boją się Wielkiego Brata AI. Są też świadomi „czerwonego oka" HAL-a, nieuchronnej symbiotycznej bliskości człowieka i inwigilującej technologii – niezbędnej, bo od niej zależy przetrwanie w pustce. Można wymieniać liczne defamiliaryzujące sceny interakcji człowiek–maszyna. I w tych trudnych relacjach wszyscy (HAL także) w pogłębiającej się nieufności i w poczuciu osamotnienia zdają się zanurzać w odmęty narastającej paranoi. Jeśli błędy algorytmów zaprojektowanych na Ziemi doprowadzały wszystkich do szaleństwa, to szaleństwem umysłu AI jest awaria jego decyzyjnej spójności. Śmierć HAL-a może być potraktowana jako kres procesu osuwania się w takie szaleństwo. Jest podobna do wyboru śmierci robota-chłopca w *A.I.* Kryzys świadomości maszyny myślącej jest jej kresem.

Taniec śmierci na Discovery

Ostateczna śmierć, fizyczna likwidacja mechanizmu superinteligencji, ma tutaj wymiar niemal szekspirowskiej tragedii, bo poprzedzona jest nieprzewidzianym przez algorytm cyklem fizycznej likwidacji załogi. Najpierw w otwartej przestrzeni, poza statkiem uśmiercony zostaje Poole, potem giną hibernanci i w końcu HAL podejmuje nieudane próby wyeliminowania Bowmana. Ten szczególny *danse macabre* wpisuje się w bezwzględny cel misji, okazuje się deterministycznym „DNA" wyprawy. Szczególnie dojmująca jest zagłada uśpionych astronautów. Nelson zwraca uwagę, że „hibernakule w kształcie trumien"[36] już wcześniej, gdy malował je Bowman, były niepokojącymi sygnaturami śmierci.

> [Z]bliżenie na maszyny hibernacyjne i ich elektroniczne ekrany wyświetlające dane o funkcjach życiowych: oddychanie, kardiogramy, elektroencefalogramy. […] Nagle linie zaczynają gwałtownie skakać, wyświetla się migotliwy komunikat: AWARIA KOMPUTERA. Linie nierówno drgają, opadają: FUNKCJE ŻYCIOWE W STANIE KRYTYCZNYM. Linie rejestrują poziom zerowy i na nim pozostają: FUNKCJE ŻYCIOWE ZAKOŃCZONE […] najbardziej przerażająca scena śmierci, jaką można sobie wyobrazić – biedni astronauci umierają statystycznie, jako linie na wykresie[37].

Wyłączenie przez HAL-a systemów podtrzymujących funkcje życiowe jest silnie oddziałującym zobrazowaniem technologicznej formy nekropolityki, tchórzliwym unikiem bezpośredniego angażowania w zabijanie, interwencją pozwalającą umknąć racjonalizacji przemocy bezpośredniej (do tej sceny Hal komentował i wyjaśniał wszystko, co czynił). To wizjonersko trafne rozpoznanie wcielenia nowoczesnych wcieleń seryjnych zbrodni. Technicyzacja mordu przypomina amerykańskie metody zapośredniczania wykonania wyroku śmierci. W filmie scena likwidacji śpiących ludzi jest symbolicznym wydaniem wyroku superinteligencji na siebie i prefiguruje analogiczną lobotomię komputera, mściwie wykonaną przez Bowmana. Kagan celnie wskazuje, że HAL zabija Poole'a i innych astronautów, pozostawiając Bowmanowi cenne wskazówki, abecadło mordowania, które prowadzi do likwidacji komputera[38]. Co znaczące, Bowman-kat, zabijając, znajduje się we wnętrzu obwodów maszyny cyfrowej, w głowie HAL-a.

[36]Nelson, 122.
[37]Kagan, 155–157.
[38]Kagan, 166.

Gdy więc dokonuje lobotomii, powraca motyw paranoi i szaleństwa AI – *signum* kresu zarazem jej świadomości i istnienia.

> Bowman zaczyna wyciągać elementy wielkości paczek papierosów z paneli sztucznego intelektu HAL-a. […] Głos HAL-a staje się proszący, jedyny raz w filmie wzbudza empatię: „Dave. Przestań… Przestań. Czy… Przestań, Dave… Przestań, […] Dave… Boję się… Boję się, Dave… Dave… Mój umysł szaleje… Czuję to… Czuję to… Mój umysł szaleje… Nie ma co do tego wątpliwości… Czuję to… Czuję to… Czuję to… Boję się". Boleśnie poetycka lobotomia HAL-a imituje naturalną śmierć, bo prowadzi do starości i w jej efekcie do brzegu drugiego dzieciństwa. Z żalem, chwiejnie HAL śpiewa: „Dzień dobry, panowie. Jestem komputerem HAL-9000… Pan Langley nauczył mnie śpiewać piosenkę. …Nazywa się *Daisy*[39].

Zanim Bowman przystąpi do tej okrutnej egzekucji, bardzo ludzki, egzystencjalny lęk HAL-a perswazyjnie obnaża egoistyczne cele przeciwdziałania konsekwencjom gniewu astronauty. W desperacji oferuje Bowmanowi pomoc we wspólnej realizacji celów misji i rozpaczliwie błaga o życie. Już w trakcie procesu wyłączania kolejnych podzespołów elektroniczny mózg wraca do swoich „kolan dzieciństwa", śpiewa: „Daisy, Daisy, daj mi prawdziwą odpowiedź / już troszkę oszalałem z miłości do ciebie…". Banalna piosenka jest nie tylko świadectwem deleksykalizacji i zdziecinnienia SI. Kubrick podsuwa widzowi bardzo mocny efekt emocjonalny, obliczony na wywołanie w dziecięcych konotacjach bezwiednego współczucia. W scenie śmierci HAL jest bliźniaczo podobny do androida-chłopca z *A.I.* Spielberga/Kubricka. Współodczuwanie jest możliwe, gdy odkrywamy, że nie było żadnej ponadludzkiej obcości superinteligencji, że Hal był ułomnym ludzkim tworem, że powtarzał jedynie za wyraźnymi tropami zbrodni dokonywanych (także we wcześniejszych sekwencjach filmu) przez swych stworzycieli: „Ostatecznie wszechwiedza HAL-a załamuje się w napadzie paranoicznej niepewności, a następnie gaśnie przy dźwiękach wolno przewijającej się taśmy, która demaskuje tego boga jako dzieło człowieka"[40].

Bibliografia i filmografia podmiotowa

Aldiss, Brian. „Super-Toys Last All Summer Long". W tegoż: *Man in His Time. The Best Science Fiction Stories*. New York: Open Road Integrated Media, 2024.

Asimov, Isaac. *Ja, robot*. Tłum. Zbigniew A. Królicki. Poznań: Rebis, 2024 [e-book].

Kubrick, Stanley. *2001: Odyseja kosmiczna*. USA, 1968.

Spielberg, Steven. *A.I. Sztuczna inteligencja*. USA, 2001.

[39]Kagan, 157.

[40]Nelson, 129.

Bibliografia przedmiotowa

Agamben, Giorgio. *Pinokio. Przygody pajacyka podwójnie skomentowane i potrójnie zilustrowane*. Tłum. Joanna Ugniewska. Warszawa: Fundacja Augusta Hrabiego Cieszkowskiego, 2024.

Bonnefon, Jean-François, Azim Shariff, Iyad Rahwan. „The Moral Psychology of AI and the Ethical Opt-Out Problem". W: *Ethics of Artificial Intelligence*, red. S. Matthew Liao, 109–126. Oxford: Oxford University Press, 2020.

Coeckelbergh, Mark. *AI Ethics*. Cambridge: The MIT Press, 2020.

Gehrke, Pat J., Ercolini G. L. „Subjected Wills: the Antihumanism of Kubrick's Later Films". W: *Depth of Field. Stanley Kubrick, Film, and the Uses of History*, red. Geoffrey Cocks, James Diedrick, Glenn Perusek, 101–121. Madison: The University of Wisconsin Press, 2006.

Kagan, Norman. *The Cinema of Stanley Kubrick*. New York: Continuum, 1989.

Kamm, F. M. „The Use and Abuse of the Trolley Problem Self-Driving Cars, Medical Treatments, and the Distribution of Harm". W: *Ethics of Artificial Intelligence*, red. S. Matthew Liao, 79–108. Oxford: Oxford University Press, 2020.

Lecercle, Jean-Jacques. *Frankenstein: mit i filozofia*. Tłum. Piotr Herbich. Warszawa: Fundacja Evviva L'Arte, 2022.

Natali, Maurizia. „«2001: A Space Odyssey» Kubrick's Allegory of Melancholia". W: *A Critical Companion to Stanley Kubrick*, red. Elsa Colombani, 249–262. Lanham: Lexington Books, 2020.

Nelson, Thomas Allen. *Kubrick. Inside a Film Artist's Maze*. Bloomington: Indiana University Press, 1982.

Phillips, Gene D., Rodney, Hill. „HAL-9000" [hasło]. W: *The Encyclopedia of Stanley Kubrick*, red. Gene D. Phillips, Rodney Hill, 138–143. New York: Facts On File, Inc., 2002.

Rasmussen, Randy. *Stanley Kubrick. Seven Films Analyzed*. Jefferson: McFarland & Company, 2001 [e-book].

Ratna, Lawrence. „Kubrick and Madness". W: *The Bloomsbury Companion to Stanley Kubrick*, red. Nathan Abrams, I.Q. Hunter, 271–280. New York: Bloomsbury Academic, 2021.

Richardson, Kathleen. „The Complexity of Otherness Anthropological Contributions to Robots and AI". W: *The Oxford Handbook of Ethics of AI*, red. Markus D. Dubber, Frank Pasquale, Sunit Das, 554–569. Oxford: Oxford University Press, 2020.

Rosenfeld, Aaron S. *Character and Dystopia. The Last Men*. New York: Routledge, 2021.

Sikora, Joshua. „The Everlasting Moment: Enchantment and Myth in A.I. and «2001: A Space Odyssey»". W: *A Critical Companion to Stanley Kubrick*, red. Elsa Colombani, 263–276. Lanham: Lexington Books, 2020.

Smith, Bryant Walker. *Ethics of Artificial Intelligence in Transport*. W: *The Oxford Handbook of Ethics of AI*, red. Markus D. Dubber, Frank Pasquale, Sunit Das, 668–683. Oxford: Oxford University Press, 2020.

Tibbetts, John C. „A.I. Artificial Intelligence". W: *The Encyclopedia of Stanley Kubrick*, red. Gene D. Phillips, Rodney Hill, 3–8. New York: Facts On File, Inc., 2002.

Wallach, Wendell, Colin Allen. *Moral Machines. Teaching Robots Right from Wrong*. Oxford: Oxford University Press, 2009).

Wallach, Wendell, Shannon Vallor. „Moral Machines. From Value Alignment to Embodied Virtue". W: *Ethics of Artificial Intelligence*, red. S. Matthew Liao, 383–412. Oxford: Oxford University Press, 2020.

Walsh, Toby. *Machines Behaving Badly. The Morality of AI*. Cheltenham: Flint, 2022 [e-book].

SŁOWA KLUCZOWE:

etyka

AI

Abstrakt:
Artykuł omawia problem regulowania relacji między ludźmi a autonomicznymi urządzeniami SI. Autor analizuje jego znaczenie dla literatury i filmu science fiction oraz współczesnej refleksji etycznej dotyczącej sztucznej inteligencji. Kontekstem są kulturowe lęki Zachodu związane z AI, czego przykładem są literackie i filmowe przedstawienia. Autor zastanawia się również nad przyszłością AI i jej potencjalnym wpływem na ludzkość.

science fiction

Stanley Kubrick

Nota o Autorce:
Rafał Szczerbakiewicz – doktor habilitowany, profesor Uniwersytetu Marii Curie-Skłodowskiej, dyrektor Szkoły Doktorskiej Nauk Humanistycznych i Sztuki, jego zainteresowania naukowe obejmują: eseistykę XX wieku, pogranicza literatury i kultury popularnej, historię muzyki popularnej, nowe media, krytykę ideologii, problematykę mitu śródziemnomorskiego w kulturze nowoczesności, ideologię kina klasycznego.