Rozłączne scenariusze katastrofalnego ryzyka SI

Kaj Sotala

1. Wprowadzenie

Praca w dziedzinie związanej z bezpieczeństwem wymaga czegoś, co zostało nazwane „mentalnością bezpieczeństwa" (Schneier, 2008): umiejętności spojrzenia na istniejący system i zaobserwowania, w jaki sposób może on zostać zagrożony przez zdeterminowanego atakującego. Podobnie praca nad bezpieczeństwem związanym z SI wymaga analogicznego sposobu myślenia, w którym ludzie aktywnie analizują, w jaki sposób coś może pójść nie tak, zamiast zakładać, że wiarygodny pomysł na wykonanie czegoś dobrze, jest wystarczający do zapewnienia bezpieczeństwa (Arbital, 2017).

Niestety scenariusze dotyczące ryzyka związanego z wyrafinowaną SI (np. Yudkowsky, 2008a, Bostrom, 2014, Sotala i Yampolskiy, 2015) nie zawsze były przedstawiane w sposób, który wyraźnie jasno akcentował potrzebę myślenia o bezpieczeństwie SI. Powszechną krytyką jest to, że choć scenariusze te zawierają wiarygodny argument, to nie jest on w żadnym wypadku nieunikniony, a odrzucenie jakiejkolwiek kluczowej przesłanki umożliwiłoby uniknięcie scenariusza[1]. Następnie przyjmuje się, że cała analiza sugerująca taki scenariusz jest fatalnie wadliwa i można ją bezpiecznie porzucić.

Trafną odpowiedzią na taką krytykę byłoby wskazanie różnych sposobów wystąpienia katastroficznego wyniku, aby się przekonać, czy argumenty za ryzykiem rzeczywiście zależą od łatwych do obalenia przesłanek. Jednak oprócz jednego znaczącego wyjątku (A. Barrett i Baum, 2017a) nie podjęto próby systematycznej analizy różnych czynników umożliwiających katastrofę w sposób, który ułatwiłby ich analizę[2].

Ten rozdział ma na celu przedstawienie szerokiego spojrzenia na różne sposoby, w jakie rozwój wyrafinowanej SI może doprowadzić do tego, że stanie się ona wystarczająco potężna, aby spowodować katastrofę. W szczególności ten rozdział ma na celu skupienie się na sposobie, w jaki różne rodzaje ryzyka są rozłączne, na jak wiele różnych sposobów coś może pójść nie tak, z których każdy może doprowadzić do katastrofy. Czyniąc to, rozdział ma na celu rozwinięcie dotychczasowych prac (A. Barrett i Baum, 2017a), które zainicjowały stosowanie ustalonych metodologii analizy ryzyka w dziedzinie bezpieczeństwa SI (A. Barrett i Baum, 2017b).

Skoncentrowano się na SI na tyle zaawansowanej, aby można było ją uważać za OSI lub ogólną sztuczną inteligencję, raczej pomijając ryzyko związane z „wąską SI", takie jak na przykład technologiczne bezrobocie (Brynjolfsson i McAfee, 2011). Należy jednak zauważyć, że niektóre z omówionych zagrożeń, w szczególności kluczowe zdolności związane z wąskimi dziedzinami zawarte w części „Inicjator MSA: kluczowe możliwości", mogą powstać na dowolnym etapie przejścia od wąskich systemów SI do superinteligencji. Celem pracy nie było zaprzeczenie lub zminimalizowanie różnych pozytywnych aspektów, które mogą również wynikać z tworzenia SI, ani sugerowanie, że nie należy kontynuować rozwoju SI. Celem było raczej umożliwienie realizacji pozytywnego potencjału SI w taki sam sposób, w jaki lepsze zrozumienie słabości związanych z bezpieczeństwem komputerowym pozwala na tworzenie bezpiecznych systemów komputerowych.

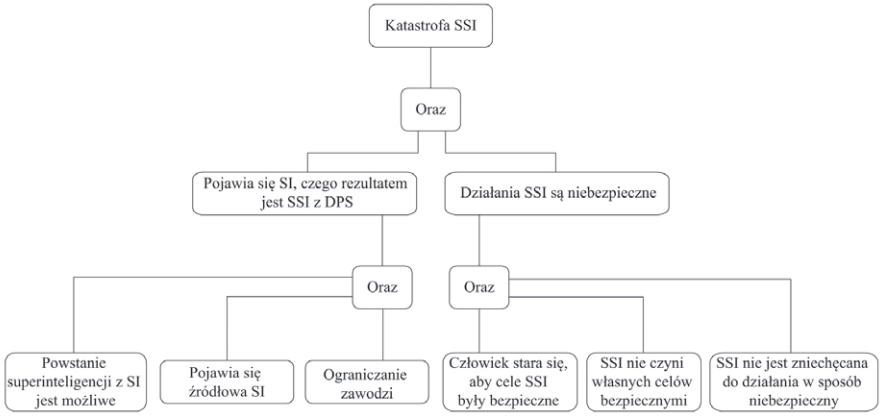
2. Inicjatorzy katastrofy

Większość argumentów za ryzykiem związanym z SI wynika z połączenia dwóch roszczeń (Yudkowsky, 2008a, Bostrom, 2014, Sotala i Yampolskiy, 2015): roszczenia dotyczącego zdolności i roszczenia dotyczącego wartości. W tym rozdziale skoncentrowano się na badaniu różnych sposobów, dzięki którym roszczenie zdolności może się spełnić. Model roszczenia wartości wykracza poza zakres tego rozdziału, aczkolwiek można zapoznać się na przykład z pracą Barretta i Bauma (2017a).

1. Roszczenie dotyczące zdolności: SI może stać się wystarczająco zdolna do potencjalnego wyrządzenia poważnych szkód dobru ludzkiemu.
2. Roszczenie dotyczące wartości: SI może działać zgodnie z wartościami, które nie są zgodne z wartościami ludzkości, powodując w ten sposób znaczne szkody.

Roszczenia te można rozpatrzeć bardziej szczegółowo. Istniejącym modelem takich roszczeń jest model SSI-PATH (A. Barrett i Baum, 2017a) (rysunek 1). SSI-PATH koncentruje się na analizie ścieżek, po których SI może doprowadzić do katastrofy, stając się superinteligentną przez rekurencyjne samodoskonalenie, przy czym ludzie nie są w stanie zapobiec tym niebezpiecznym działaniom.

Model SSI-PATH wykorzystuje konwencje schematu błędów, w których niepożądanym zdarzeniem (katastrofą SI) jest węzeł górny, po którym następują dwa węzły mogące uaktywnić górny węzeł, gdyby oba były prawdziwe. Są to węzły „Działania SSI są niebezpieczne", które odpowiadają roszczeniu wartości oraz „SI rozwija się, powstaje [Sztuczna SuperInteligencja] z [Decydująca Przewaga Strategiczna]", co odpowiada określonej formie roszczenia zdolności. W tym rozdziale rozwinięto SSI-PATH przez rozważenie bardziej ogólnych form roszczenia zdolności.



Rys. 1. Górne warstwy modelu SSI-PATH. (Na podstawie: Barrett i Baum, *Journal of Experimental & Theoretical Artificial Intelligence: JETAI* 29 nr 2, 2017a: 397–414). Warstwy te zostały zaprojektowane jako drzewo błędów, obrazujące różne warunki, które muszą zostać spełnione, aby nastąpiła katastrofa związana z SSI. Według schematu katastrofa SSI zdarza się wtedy, gdy: (1) SI rozwija się, czego efektem jest SSI z DPS oraz (2) działania SSI są niebezpieczne, powodując katastrofalne użycie DPS. Dolne węzły wskazują trzy przypadki, które muszą być prawdziwe, aby SI mogła się rozwinąć, oraz kolejne trzy przypadki, które muszą zaistnieć, aby działania SSI były niebezpieczne. Pełny model zawiera dodatkowe warstwy, które nie zostały pokazane na rysunku. Więcej szczegółów można znaleźć w Barrett i Baum (2017a)

Roszczenie zdolności jest często formułowane jako możliwość osiągnięcia przez SI decydującej przewagi strategicznej (DPS). Pojęcie DPS było przyjmowane domyślnie w wielu wcześniejszych pracach, a koncepcja ta została po raz pierwszy wyraźnie zdefiniowana przez Bostroma (2014, s. 78) jako „poziom technologicznych i innych korzyści wystarczających, aby umożliwić [SI] osiągnięcie pełnej dominacji nad światem".

Jednakże założenie, że SI osiągnie DPS wydaje się niepotrzebnie silną formą roszczenia zdolności, ponieważ SI może spowodować katastrofę niezależnie od niego. Rozważmy na przykład scenariusz, w którym SI rozpoczyna atak obliczony na zniszczenie ludzkiej cywilizacji. Jeśli SI udałoby się zniszczyć ludzkość lub jej dużą część, ale w rezultacie sama SI również zostałaby zniszczona, to nie liczyłoby się to jako DPS, jak pierwotnie zdefiniowano. Trudno jednak zaprzeczyć, że wynik taki należy jednak uznać za katastrofę.

Z tego powodu rozdział ten koncentruje się na sytuacjach, w których SI osiąga przynajmniej znaczną przewagę strategiczną (ZPS), którą określamy jako „poziom technologiczny i inne korzyści wystarczające, aby stanowić katastrofalne ryzyko dla społeczeństwa ludzkiego". Katastrofalne ryzyko to takie, które może spowodować poważne szkody dla dobrobytu ludzi w skali globalnej i spowodować 10 milionów lub więcej ofiar śmiertelnych (Bostrom i Ćirković, 2008).

Oprócz oczywistych przyczyn chęci uniknięcia katastroficznego ryzyka spowodowanego przez SI, zauważamy, że zniszczenia na szeroką skalę mogą przyczynić się do globalnych zawirowań (Bostrom i in., 2016), sytuacji, w której istniejące instytucje byłyby zagrożone, a koordynacja i długoterminowe planowanie stałyby się także trudniejsze. Globalne turbulencje mogłyby następnie przyczynić się do kolejnego niekontrolowanego projektu SI, który zawiódłby jeszcze bardziej katastrofalnie i spowodowałby jeszcze większe szkody. Zatem to, co pierwotnie było jedynie katastroficznym ryzykiem, może przyczynić się do dalszego rozwoju ryzyka egzystencjalnego (Bostrom, 2002, 2013, Sotala i Gloor, 2017).

Znaczna część istniejącej literatury na temat bezpieczeństwa SI koncentruje się na badaniu scenariuszy, w których SI osiąga DPS, oraz na analizie warunków do tego prowadzących. Jest to pod

wieloma względami rozsądna strategia, ponieważ jeśli bylibyśmy w stanie poradzić sobie z SI, która mogłaby osiągnąć DPS, to najprawdopodobniej bylibyśmy również w stanie poradzić sobie z SI, która mogłaby osiągnąć ZPS, zakładając, że silniejsza SI jest konserwatywnym założeniem (Yudkowsky, 2001). Jednak ta strategia ma tę wadę, że może sprawiać wrażenie, że znaczna część analizy bezpieczeństwa SI jest nieistotna, jeśli okaże się, że możliwość uzyskania DPS przez SI jest wyjątkowo nieprawdopodobne. Niektóre mechanizmy obronne mogą być również wystarczające, aby uniemożliwić SI uzyskanie DPS, ale nie są wystarczające, aby zapobiec uzyskaniu ZPS.

3. Kiedy zostaną podjęte działania przeciwko przewadze strategicznej?

SI, która jest w stanie wyrządzić znaczne szkody dobrobytowi ludzi, jest szczególnie groźna, gdy ma do tego motywację[3]. Istnieje również możliwość, że SI zamierzająca współpracować z ludzkością może spowodować szkody przez przypadek, wykracza to jednak poza zakres niniejszej analizy. Pomimo że pełna analiza roszczenia wartości wykracza poza zakres tego rozdziału, to nie można jej całkowicie odseparować od roszczenia zdolności, ponieważ wartości SI wpływają również na próg zdolności, przy którym racjonalne staje się dla niej działanie przeciwko ludzkości. Jak omówiono, niektóre wartości i sytuacje zwiększają prawdopodobieństwo podjęcia wrogich działań przez SI, nawet jeśli ma niewielkie możliwości.

Dwa główne powody, dla których SI może podjąć działania powodujące szkody dla ludzkości, to:

- Szkodziłaby ludzkości w dążeniu do celu, który neguje ludzkie dobro, na przykład przez rozebranie ludzkich miast w poszukiwaniu surowców. „SI ani cię nie nienawidzi, ani cię nie kocha, ale jesteś zbudowany z atomów, które może wykorzystać do czegoś innego" (Yudkowsky, 2008a, s. 333).
- Może oczekiwać, że ludzie podejmą działania przeciwko niej, co uniemożliwiłoby jej osiągnięcie celów, dlatego może podjąć działania w ich obronie, przeprowadzając atak zapobiegawczy. Byłoby to racjonalnym działaniem, ponieważ pozwoliłoby SI faktycznie zrealizować jej cele (Omohundro, 2007, 2008). Mogłoby się tak zdarzyć nawet wtedy, gdyby SI miała cel uwzględniający elementy ludzkiego dobrobytu, jeśli tylko SI znalazłaby powody, by sądzić, że ludzie mimo wszystko sprzeciwią się realizacji tego celu[4].

Dokładne cele, jakie ma SI, wpływają na poziom zdolności, których potrzebuje do tego, by wrogie ludziom działania uznać za racjonalną strategię. SI, która troszczy się głównie o jakiś mocno sprecyzowany cel, może chcieć zniszczyć ludzką cywilizację, aby mieć pewność, że potencjalne zagrożenie tego celu zostanie wyeliminowane. Dzięki temu SI mogłaby kontynuować realizację swojego celu bez przeszkód. Jednak SI, która zostałaby zaprogramowana tak, aby maksymalizować coś takiego jak „szczęście obecnie żyjących ludzi", mogłaby być znacznie mniej skłonna zaryzykować znaczną liczbę ofiar śmiertelnych[5]. Zmusiłoby to ją do skupienia się na mniej niszczycielskich metodach przejmowania potencjalnie wymagających bardziej wyrafinowanych umiejętności.

W rezultacie wartości SI określają poziom zdolności, jaki musi mieć, aby wrogie działanie było wykonalną strategią. W uproszczonym modelu (Shulman, 2010) SI uważająca, że zainicjowanie agresywnych działań ma prawdopodobieństwo odniesienia sukcesu P oraz oczekiwaną użyteczność UE(Sukces), jeśli się powiedzie, UE(Niepowodzenie), jeśli się nie powiedzie, i UE(Współpraca), jeśli zaprzestanie agresji i nadal będzie współpracować, racjonalnie zainicjuje agresję, jeśli:

$$P \times UE(Sukces) + (1 - P) \times \times UE(Niepowodzenie) > UE(Współpraca).$$

Można to uznać za sugestię, że SI przeprowadziłaby atak przede wszystkim wtedy, gdyby miała DPS lub myślała, że może ją zdobyć, a tym samym ustanowić dominację nad ludźmi. Jednak nawet SI z tylko ZSA może podjąć wrogie działania, stosując środki, takie jak wymuszenie i groźby wyrządzenia bardziej ograniczonych szkód, w celu zdobycia większej ilości zasobów lub skierowania świata w bardziej sprzyjającym kierunku.

Między innymi może się to zdarzyć:

- Jeśli SI nabrałaby tempa we własnym rozwoju, uzyskując tym samym zdolność do autonomicznego działania i wierzyła, że nie można jej wyśledzić (zobacz części od „Wyzwanie techniczne" do „Dobrowolne uwolnienie z desperacji", gdzie omówiono sposoby, w jakie SI może uzyskać swobodę lub zostać dobrowolnie uwolniona przez jej twórców).
- Gdyby SI miała sojuszników, którzy chroniliby ją przed odwetem (zobacz część „Inicjator ZPS: kluczowe zdolności", gdzie zamieszczono informacje na temat umiejętności manipulacji społecznych oraz część „Wyzwanie społeczne", aby dowiedzieć się, w jaki sposób autonomiczna SI może pozyskać ludzkich sojuszników).
- Jeśli SI kontrolowałaby ludzką organizację, której nie można zaatakować bez olbrzymich, postronnych zniszczeń (zobacz części „Inicjator DPS/ZPS: SI stopniowo przejmuje władzę" oraz „SI pozostaje ograniczona, jednak ostatecznie przejmuje kontrolę", gdzie opisano przejęcie kontroli nad ludzką organizacją).
- Gdyby istniały już silniejsze systemy SI podejmujące działania, a SI uznałaby siebie za zbyt mało wartą odwetu (zobacz część „Uwagi na temat pojedynczej i licznej SI", gdzie omówiono liczną SI).

Niezależnie od skali agresji na zachowanie SI wpływają również różne inne czynniki sytuacyjne. Na przykład SI może nie być skłonna do powodowania szkód, ponieważ mogłaby pomyśleć, że spowoduje to zbyt wiele szkód ubocznych wobec rzeczy, które ceni, ponieważ nie uważałaby się za zdolną do przetrwania wynikającego z jej działań odwetu lub ponieważ oszacowałaby, że wynikające z takiej agresji szkody w infrastrukturze pozbawiłyby ją zasobów (takich jak elektryczność), potrzebnych do jej przetrwania.

Ataki różnią się także zakresem, w jakim można je selektywnie zastosować. Tradycyjna broń palna może być wycelowana wybiórczo, podczas

Tabela 1. Czynniki zwiększające prawdopodobieństwo, że SI będzie działać wbrew ludzkim interesom

Czynniki zwiększające prawdopodobieństwo ataku	Czynniki zmniejszające prawdopodobieństwo ataku
Wysokie subiektywne prawdopodobieństwo sukcesu lub oczekiwanie niewielkiego odwetu w przypadku niepowodzenia	Niskie subiektywne prawdopodobieństwo sukcesu i oczekiwanie późniejszego odwetu
Ochrona przed odwetem przez sojuszników lub kontrolą potężnej organizacji	Niechęć do utraty sojuszników lub kontroli nad organizacją z powodu odwetu
Cel, który nie uwzględnia obecnie żyjących ludzi	Cel, który ceni obecnie żyjących ludzi[a]
Możliwość ukrycia się przed odwetem	Brak możliwości przemieszczania lub ukrycia się przed działaniami odwetowymi
Możliwość przeprowadzania ataków, które pozwalają uniknąć uszkodzenia kluczowej infrastruktury lub innych cennych celów	Możliwość przeprowadzania tylko masowych ataków
Ryzyko zbliżającego się zniszczenia	Łatwo satysfakcjonujące lub trywialne wartości
Zaawansowana zdolność do budowy lub przebudowy infrastruktury	Cele o niskim wpływie
Wysoki poziom istniejącej automatyzacji zmniejszający zależność od pracowników	
Istnienie innych SI, które mogą pierwsze uzyskać DPS	
Irracjonalność	Irracjonalność

[a] W zależności od stopnia, w jakim obecnie żyjący ludzie są cenieni: „schwytaj, nie zabijaj" może być implikowane przez niektóre pozornie korzystne cele (Williamson, 1947), aczkolwiek nawet cele, które „tylko" zabraniają ludzkiej śmierci są trudniejsze do osiągnięcia niż cele, które pozwalają na więcej szkód ubocznych.

gdy pandemie potencjalnie zagrażają wszystkim przedstawicielom gatunku. W zakresie, w jakim SI musi polegać na ludzkiej gospodarce, która wytwarza zasoby potrzebne do przetrwania SI, ataki zagrażające gospodarce również zagrażają zasobom SI. Zasoby te są w pewnym sensie dzielone pomiędzy SI i ludzkość, tak więc wszelkie ataki, które powodują masowe uszkodzenia tych zasobów, są niebezpieczne dla obu stron. Im bardziej SI może projektować ataki selektywnie pozbawiające przeciwników zasobów, tym niższy jest próg ich wykorzystania. Bardziej zaawansowane możliwości przebudowy infrastruktury pozwoliłyby SI na przeprowadzenie bardziej masowego ataku. SI, która była w stanie zbudować bardziej zaawansowaną infrastrukturę niż obecnie istniejąca, mogłaby zlekceważyć uszkodzenia obecnej infrastruktury, jeśli i tak planowałaby zburzyć jej większość.

Bilans tych kalkulacji mógłby zostać przesunięty, gdyby SI myślała, że grozi jej zniszczenie przez ludzi, nawet gdyby współpracowała (obniżając oczekiwaną użyteczność współpracy). Samozachowawczość jest instrumentalnym celem wielu różnych wartości, ponieważ istniejący agent jest bardziej zdolny do promowania większości wartości niż agent, który nie istnieje (Omohundro, 2007, 2008, Bostrom, 2012)[6]. SI, która znalazłaby się w bezpośrednim niebezpieczeństwie zniszczenia, mogłaby racjonalnie zainicjować kontratak, ryzykując nawet duże zniszczenia, o ile oszacowałaby, że oczekiwana wartość scenariusza, w którym kontratak umożliwiłby jej przetrwanie i promowanie jej wartości, przewyższałaby szkody spowodowane przez taki kontratak. Byłoby to szczególnie przekonującym czynnikiem motywującym, gdyby SI miała idiosynkratyczne wartości, które jej zdaniem z małym prawdopodobieństwem byłyby promowane przez innych agentów. Gdyby istniało wiele projektów SI i SI uwierzyłaby, że jeden z innych projektów może pierwszy uzyskać DPS, to byłby to wystarczający powód, by zaryzykować wcześniejszy atak (zobacz część „Uwagi na temat pojedynczej i licznej SI", gdzie zawarto opis licznej SI). Pojawiły się również propozycje zaprojektowania wartości SI w sposób, który wyraźnie obniża wartość wrogiego działania[7].

W powyższej analizie założono, że SI wybiera swoje działania racjonalnie.

Irracjonalność może wydawać się czymś, co uniemożliwiłoby SI uzyskanie bardzo dużych zdolności, jednak, podobnie jak ludzie, SI mogłaby być pod niektórymi względami racjonalna, a pod innymi nieracjonalna. Dla SI może być również racjonalne podjęcie działań pozornie nieracjonalnych, na przykład poprzez irracjonalne ignorowanie zagrożeń, tak aby inni uważali próby zagrażania jej za mniej opłacalne (Parfit, 1984, część 5). Główną kwestią wynikającą z potencjalnej nieracjonalności jest to, że nie można po prostu polegać na tym, że SI nie spowoduje uszkodzeń, nawet jeśli byłby to racjonalny sposób jej zachowania. Oczywiście nieracjonalność może również spowodować, że SI uniknie wyrządzania szkód w sytuacji, gdyby było to racjonalne (tabela 1).

4. Inicjatorzy katastroficznych zdolności

W tej części rozważono cztery ogólne scenariusze, według których SI mogłaby uzyskać DPS lub ZPS: scenariusze indywidualnego wejścia w życie z jego trzema głównymi podtypami, scenariusze zbiorowego wejścia w życie, scenariusze stopniowego przejęcia kontroli przez systemy SI oraz scenariusze, kiedy SI staje się wystarczająco dobra w niektórych kluczowych możliwościach i uzyskuje ZPS lub DPS.

Na prawdopodobieństwo sukcesu lub porażki każdego z tych scenariuszy wpływa również to, jaką zdolnością do współpracy wykażą się ludzie. Pomimo że możliwe są scenariusze, w których SI staje się całkowicie samodzielna i musi uniemożliwić twórcom jej wyłączenie, to istnieje również wiele możliwych scenariuszy omówionych w części „SI uzyskuje zdolność do samodzielnego działania", w których SI uzyskuje częściową lub pełną współpracę swoich twórców, przynajmniej do pewnego momentu. Taki rozwój wydarzeń wpłynąłby na prawdopodobieństwo spełnienia się każdego z poniższych scenariuszy. Scenariusz, w którym prototypowa SI musi unikać jej zamknięcia przez programistów, różni się bardzo od scenariusza, w którym programiści są pewni, że SI jest bezpieczna i dobrowolnie pomagają jej gwałtownie się rozwinąć, szczególnie jeśli mają do dyspozycji zasoby dużej korporacji lub państwa.

Inicjatorzy DPS: scenariusze wejścia w życie

„Odejście" (Bugaj i Goertzel, 2007) to proces, w którym SI staje się znacznie bardziej zdolna niż ludzkość. W przypadku łagodnego wejścia w życie dzieje się to stopniowo w czasie, co pozwala na ciągłą interakcję człowieka, podczas gdy w przypadku gwałtownego wejścia w życie po przekroczeniu pewnego etapu SI bardzo szybko zwiększa swoje zdolności, wyrywając się ze skutecznej kontroli człowieka.

Warto zauważyć, że w przypadku gwałtownego wejścia w życie nie zakłada się, że SI stanie się bardzo zdolna natychmiast po stworzeniu (jednak moment jej utworzenia jest określony). Scenariusz gwałtownego wejścia w życie może obejmować wydłużony okres stopniowego rozwoju, aż do osiągnięcia pewnego kluczowego poziomu zdolności, od którego SI gwałtownie się rozwija.

Wiele wcześniejszych dyskusji (np. Yudkowsky, 2008a, Bostrom, 2014, Sotala, 2017) koncentrowało się na analizie możliwości gwałtownego wejścia w życie. Chociaż nie jest to jedyny możliwy scenariusz, w którym SI może stać się zdolna, to jest to scenariusz, który pozostawia najmniej możliwości przeciwdziałania złemu rozwojowi zdarzeń.

Mając na uwadze, że nadmierne skupienie się na scenariuszach gwałtownego wejścia w życie może zamaskować fakt, że nie jest on konieczny do tego, aby SI mogła uzyskać ZPS lub DPS, najpierw rozważymy scenariusze gwałtownego wejścia w życie, a następnie inne czynniki inicjujące.

Inicjator DPS: indywidualne wejście w życie

„Indywidualne wejście w życie" to takie, w którym pojedyncza SI staje się tak potężna, że całkowicie dominuje ludzkość. W literaturze zaproponowano trzy ogólne ścieżki prowadzące do takiego scenariusza: nadwyżka sprzętowa („więcej SI"), eksplozja prędkości („szybsza SI") i eksplozja inteligencji („inteligentniejsza SI") (Sotala i Yampolskiy, 2015). Bostrom (2014) omówił je w kategoriach odpowiednio superinteligencji kolektywnej, szybkiej superinteligencji i jakościowej superinteligencji. Należy zauważyć, że ścieżki te nie wykluczają się wzajemnie i wręcz przeciwnie, każda z nich może przyczynić się do rozwoju drugiej.

Nadwyżka sprzętowa

W scenariuszu nadwyżki sprzętowej (Yudkowsky, 2008b, Shulman i Sandberg, 2010) sprzęt rozwija się szybciej niż oprogramowanie, dzięki czemu mogą zaistnieć komputery o większej mocy obliczeniowej niż ludzki mózg, jednak bez możliwości efektywnego wykorzystania całej tej mocy. Gdyby jednak ktoś opracował algorytm ogólnej inteligencji mogącej efektywnie wykorzystać taki sprzęt, to nagle mogłoby pojawić się mnóstwo taniego sprzętu, który mógłby zostać wykorzystany do uruchamiania tysięcy lub milionów kopii SI. Taka liczna SI mogłaby, ale i nie musiałaby być superinteligentna, jednak sama ich liczba pozwoliłaby SI na prowadzenie skoordynowanych operacji na masową skalę. Gdyby pojedyncza SI wykorzystała ten potencjał do wytworzenia dużej liczby swoich kopii lub subagentów, to umożliwiłoby to jej indywidualne wejście w życie[8]. W przeciwnym razie stanowiłoby to zbiorowe wejście w życie, jak omówiono to poniżej.

Nadwyżka sprzętowa może się faktycznie wydarzyć, nawet jeśli SI byłaby początkowo ograniczona sprzętowo: pierwsze jednostki SI mogą wymagać dużej ilości sprzętu, jednak dalsze optymalizacje szybko mogą obniżyć wymagania sprzętowe. Patrząc na ostatnie postępy w rozwoju SI, początkowe podejście do nauki gier Atari 2600 (Mnih i in., 2015) wykorzystywało specjalistyczny sprzęt w postaci GPU, jednak dopiero rok później wydano alternatywne podejście, w którym wykorzystano standardowy procesor i osiągnięto lepsze wyniki przy użyciu krótszego czasu uczenia (Mnih i in., 2016). Oprócz sugestii, że optymalizacje oprogramowania mogą szybko zwiększyć liczbę możliwych do uruchomienia kopii SI, to także fakt poprawy szybkości i wydajności podkreśla możliwość

wystąpienia scenariusza nadwyżki sprzętowej, który jednocześnie przyczynia się do możliwości wystąpienia scenariuszy eksplozji prędkości i eksplozji inteligencji, jak omówiono poniżej.

Eksplozja prędkości

W scenariuszu eksplozji prędkości (Solomonoff, 1985, Yudkowsky, 1996, Chalmers, 2010) inteligentne maszyny projektują coraz szybsze maszyny. Nadwyżka sprzętowa może się przyczynić do eksplozji prędkości, nie jest jednak ona warunkiem koniecznym. SI działająca w tempie człowieka mogłaby opracować sprzęt drugiej generacji, na którym mogłaby działać w znacznie szybszym tempie niż ludzkie myśli. Opracowanie sprzętu kolejnej, trzeciej generacji wymagałoby zatem krótszego czasu i umożliwiłoby SI działać jeszcze szybciej niż poprzednia generacja i tak dalej. W pewnym momencie proces dotarłby do fizycznych granic i zatrzymałby się, jednak do tego czasu sztuczna inteligencja mogłaby wykonać większość zadań w znacznie szybszym tempie niż ludzie, osiągając w ten sposób dominację. Zasadniczo można to również osiągnąć za pomocą ulepszonego oprogramowania, jak to wcześniej omówiono.

Stopień, w jakim SI potrzebuje ludzi do wyprodukowania lepszego sprzętu, ogranicza tempo eksplozji prędkości, tak więc szybka eksplozja prędkości wymaga zdolności do automatyzacji dużej części procesu produkcji sprzętu. Jednak ten rodzaj automatyzacji może zostać osiągnięty do czasu opracowania SI. Im większa automatyzacja, tym szybciej może nastąpić zdobycie dominacji przez SI.

Jeśli poziom bezpieczeństwa sprzętu byłby dobry, to scenariusze eksplozji szybkości, w których SI włamuje się do systemów produkcyjnych i przejmuje nad nimi kontrolę, stają się mniej prawdopodobne. Z drugiej strony istnieją możliwe ścieżki, omówione w części „SI uzyskuje zdolność do samodzielnego działania", w których SI uzyskuje prawowitą kontrolę nad różnymi zasobami. Zapewnienie odpowiedniej kontroli bezpieczeństwa zautomatyzowanym fabrykom nie byłoby pomocne, jeśli byłyby one kierowane przez SI lub jeśli SI mogłaby uzyskać do nich dostęp na otwartym rynku i miałaby na ten cel wystarczającą ilość środków.

Eksplozja prędkości może również przyczynić się do zaistnienia nadwyżki sprzętowej i eksplozji inteligencji, umożliwiając znalezienie bardziej wydajnych lub w inny sposób lepszych algorytmów w krótszym czasie.

Eksplozja inteligencji

Podczas eksplozji inteligencji (Dobry, 1965, Chalmers, 2010, Bostrom, 2014) SI wymyśla, jak stworzyć jakościowo inteligentniejszą SI i następnie ta inteligentniejsza SI wykorzystuje swoją zwiększoną inteligencję do stworzenia jeszcze bardziej inteligentnej SI i tak dalej. W ten sposób ludzka inteligencja pozostałaby daleko w tyle, a maszyny osiągnęłyby dominację.

W wielu dziedzinach istnieją granice przewidywania na podstawie eksplozji kombinatorycznych, które wynikają z próby prognozowania coraz bardziej w przyszłość. Na przykład w modelowaniu prognozy pogody można uzyskać dostęp tylko do ograniczonej liczby wstępnych obserwacji w odniesieniu do liczby stopni swobody systemu przewidywania pogody (Buizza, 2002). Jednak, nawet jeśli superinteligentna SI nie byłaby w stanie dokładnie przewidzieć każdego przyszłego zdarzenia, to nadal mogłaby zareagować na to zdarzenie i przewidzieć jego prawdopodobne konsekwencje lepiej niż ludzie. Tetlock i Gardner (2015) dokonali przeglądu i omówili zdolność niektórych ludzkich prognostów („superprognostów") do przewidywania wydarzeń na świecie ze znaczną dokładnością. Na temat nieprzewidywalnych wydarzeń zwanych „czarnymi łabędziami" (Taleb, 2007) Tetlock i Gardner (2015, Kindle lok. 3614) piszą:

> Możemy nie mieć żadnych dowodów na to, że superprogności mogą przewidzieć wydarzenia takie jak te z 11 września 2001.

> Istnieje jednak cały szereg dowodów na to, że mogą prognozować pytania, takie jak: czy Stany Zjednoczone zagrożą działaniami wojskowymi, jeśli talibowie nie przekażą Osamy bin Ladena? Czy talibowie zgodzą się na to? Czy bin Laden ucieknie z Afganistanu przed inwazją? W zakresie, w jakim takie prognozy mogą przewidzieć konsekwencje wydarzeń podobnych do tych z 11 września, a konsekwencje takie sprawiają, że czarny łabędź jest tym, czym jest, to możemy przewidzieć wystąpienie czarnych łabędzi.

Sotala (2017), na podstawie przeglądu literatury na temat ludzkiej wiedzy i inteligencji, stwierdza, że u ludzi wiedza specjalistyczna opiera się na rozwijaniu wyobrażeń mentalnych, które pozwalają ekspertom zrozumieć różne sytuacje i albo natychmiast poznać odpowiednie działania w danej sytuacji, albo przeprowadzić mentalną symulację tego, jak może się rozwinąć taka sytuacja i jaka powinna być na nią reakcja. Taką wiedzę specjalistyczną zapewnia połączenie dwóch umiejętności: rozpoznawania wzorców i symulacji mentalnej.

Sotala (2017) twierdzi, że SI mogłaby usprawnić obie umiejętności. Zdolność nadludzkiej symulacji mentalnej można osiągnąć przez połączenie wykonywania bardziej złożonych symulacji z uwzględnieniem większej liczby czynników, a także poprzez wykorzystanie kilku strumieni uwagi, które mogłyby badać wiele alternatywnych metod równolegle, jednocześnie analizując wiele różnych perspektyw i czynników przyczynowych. Przeprowadzanie dokładnych symulacji mentalnych wymagałoby również dobrej reprezentacji mentalnej w celu utworzenia podstawowych elementów składowych symulacji. Wśród ludzi istnieją różnice poznawcze, które pozwalają niektórym ludziom uczyć się i uzyskiwać dokładne reprezentacje mentalne szybciej niż inni i wydaje się, że sprowadzają się one do takich czynników, jak pojemność pamięci roboczej, kontrola uwagi i pamięć długoterminowa. Czynniki te można udoskonalić przez połączenie ulepszeń sprzętowych i teoretycznej informatyki. Wydaje się, że u ludzi ulepszenie inteligencji zapewnia dodatkowe korzyści w całym udokumentowanym zakresie różnic inteligencji i wydaje się prawdopodobne, że różne ograniczenia ewolucyjne przyczyniły się do ograniczenia rozwoju ludzkiej inteligencji znacznie poniżej teoretycznego maksimum. W odniesieniu do ograniczeń prognozowania wynikających z wewnętrznej niepewności świata, Sotala (2017, s. 12) uznaje istnienie takich ograniczeń, jednak twierdzi, że:

> wygląda na to, że chociaż system SI od samego początku nie byłby w stanie stworzyć jednego superplanu podboju świata, to wciąż miałby nadludzką zdolność adaptacji i uczenia się na podstawie zmieniających się i nowatorskich sytuacji oraz reagowania na nie szybciej niż ludzcy przeciwnicy. Analogicznie, eksperci grający w większość gier nie są w stanie obliczyć zwycięskiej strategii już od pierwszego ruchu, jednak nadal mogą reagować i dostosowywać się do zmieniającej się sytuacji gry lepiej niż nowicjusz, co pozwala im wygrać.

Eksplozja inteligencji może również przyczynić się do wystąpienia eksplozji prędkości i nadwyżki sprzętowej, jeśli zwiększona inteligencja SI umożliwiłaby jej znalezienie algorytmów, które byłyby najbardziej wydajne pod względem możliwości uruchomienia większej liczby systemów SI z tym samym sprzętem (nadwyżka sprzętowa) lub możliwości szybszego uruchomienia (eksplozja prędkości).

Inicjator DPS: zbiorowe wejście w życie handlującej SI

Vinding (2016), a także Hanson i Yudkowsky (2013) argumentują, że duża część pozornie indywidualnej ludzkiej inteligencji w rzeczywistości opiera się na możliwości korzystania z rozproszonych zasobów całej ludzkości, zarówno tych materialnych, jak i poznawczych. Z tego powodu błędem może być skupienie się na punkcie, w którym SI osiągają inteligencję na poziomie ludzkim, ponieważ inteligencja zbiorowa jest ważniejsza niż inteligencja indywidualna. Najłatwiejszym dla SI sposobem na osiągnięcie poziomu zdolności porównywalnego z ludzkim byłaby współpraca ze społeczeństwem ludzkim i pokojowe wykorzystanie jego zasobów.

Hall (2008) podobnie zauważa, że nawet gdy pojedyncza SI dokona samodoskonalenia, na przykład opracowując lepsze modele kognitywistyki w celu ulepszenia swojego oprogramowania, to reszta gospodarki również będzie rozwijać takie lepsze modele. Z tego powodu dla SI korzystniejsze jest skupienie się na ulepszaniu wszystkiego, w czym jest najlepsza, i kontynuowanie handlu z resztą gospodarki oraz kupowanie tych rzeczy, w których reszta gospodarki jest lepsza od niej.

Jednak Hall zauważa, że nadal może nastąpić gwałtowne wejście w życie SI w momencie, gdy wystarczająca liczba kopii SI zostanie połączona w sieć. SI, która myśli szybciej niż ludzie, może się ze sobą komunikować i dzielić się spostrzeżeniami znacznie szybciej, niż może to robić z ludźmi. W rezultacie dla SI zawsze byłoby lepiej handlować i współpracować z innymi SI niż z ludźmi. Wielkość gospodarki SI może rosnąć dość szybko, a Hall (s. 464) sugeruje scenariusz: „od […] 30 000 równoważników ludzkich na początku do około 5 miliardów równoważników ludzkich dekadę później". Nawet więc jeśli żadna pojedyncza SI nie mogłaby sama osiągnąć DPS, to wspólna społeczność SI mogłaby ją osiągnąć, ponieważ taka społeczność rozwinęła się tak, aby była zdolna do wszystkiego, co ludzie byli w stanie osiągnąć[9].

Inicjator DPS/ZPS: SI stopniowo przejmuje władzę

Historycznym trendem było zautomatyzowanie wszystkiego, co można było zautomatyzować zarówno w celu zmniejszenia kosztów, jak i dlatego, że maszyny mogą robić rzeczy lepiej niż ludzie. Każda firma mogłaby potencjalnie lepiej funkcjonować, gdyby była prowadzona przez umysł, który został specjalnie zaprojektowany do prowadzenia danej firmy, włącznie z zastąpieniem wszystkich pracowników jednym lub większą liczbą takich umysłów. SI może myśleć szybciej i mądrzej, radzić sobie z większą ilością informacji naraz i pracować w jednym celu, zamiast osłabiać swoją efektywność przez politykę biurową, która nęka każdą dużą organizację. Niektóre szacunki już sugerują, że połowa

zadań, za które ludzie są wynagradzani jest podatna na automatyzację przy użyciu technik współczesnego uczenia maszynowego i robotyki, nawet bez wprowadzania SI z ogólną inteligencją (Frey i Osborne, 2013, Manyika i in., 2017).

Tendencja do automatyzacji trwała przez całą historię, nie wykazuje żadnych oznak słabnięcia i nieodłącznie wiąże się z udzielaniem systemom SI dowolnych, potrzebnych możliwości, tak aby mogła lepiej zarządzać firmą. Istnieje ryzyko, że systemy SI, które początkowo były proste i miały ograniczoną inteligencję, będą stopniowo zdobywały coraz większą moc i odpowiedzialność, w miarę jak będą się uczyły i będą ulepszane, dopóki znaczna część społeczeństwa nie znajdzie się pod kontrolą SI.

Inicjator ZPS: kluczowe możliwości

W przypadku omawiania ZPS kluczową kwestią jest próg zdolności wystarczający do zadania katastroficznych szkód. SI mogłaby być katastrofalnym ryzykiem, gdyby jej zdolności ofensywne w niektórych kluczowych dziedzinach były wystarczające do pokonania istniejącej obrony.

Jak krótko omówiono to w części „Kiedy zostaną podjęte działania przeciwko przewadze strategicznej?", zakładając, że SI byłaby racjonalna, to wybór spowodowania takich szkód wymagałby rozsądnego motywu. Jednak podobnie jak w przypadku ludzi, może istnieć szereg motywów, które uczynią rozsądną strategią wrogie działania, takie jak wymuszenie, chęć pomocy sojusznikowi lub atak uprzedzający przeciwko innej SI lub grupie mogącej uzyskać DPS. W zależności od celów i od tego, czy SI miałaby sojuszników, przeprowadzenie ataku możliwego ze względu na kluczowe zdolności może wymagać posiadania dodatkowych możliwości, takich jak odbudowa po zniszczeniu kluczowej infrastruktury.

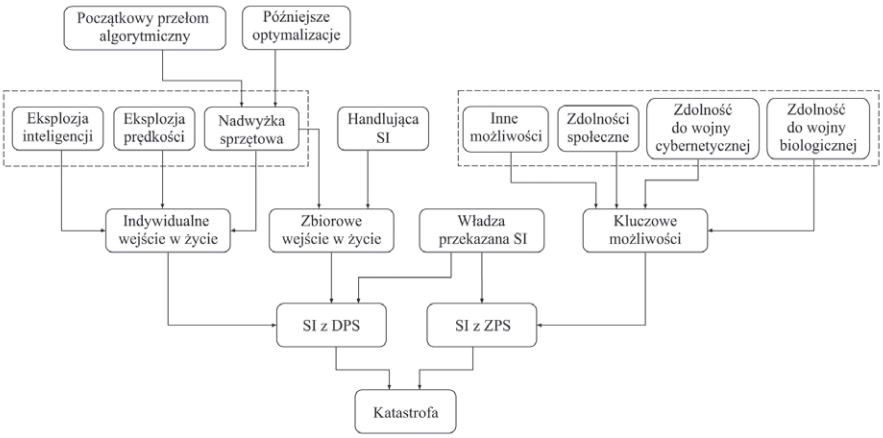
Należy zauważyć, że powodowanie katastrofalnych uszkodzeń prawdopodobnie nawet nie wymaga nadludzkich zdolności (Torres, 2016, 2017, rozdz. 4). Na przykład wydaje się możliwe, że wystarczająco zdeterminowany ludzki napastnik mógłby obecnie spowodować poważne szkody w społeczeństwie przez wojnę elektroniczną. Chociaż nie odnotowano jeszcze cyberataków, które mogłyby bezpośrednio przyczynić się do śmierci, to kilka z nich spowodowało szkody fizyczne lub zakłócenia w działaniu służb ratunkowych. W maju 2017 roku ogłoszono, że robak ransomware „WannaCry" zainfekował ponad 230 000 komputerów w ponad 150 krajach (Ehrenfeld, 2017), powodując zakłócenie działania kluczowych usług, takich jak opieka zdrowotna (Gayle i in., 2017). W 2016 roku ogłoszono, że trzy podstacje w ukraińskiej sieci energetycznej zostały odłączone w wyniku ataku

złośliwego oprogramowania, pozostawiając bez prądu około połowę domów w regionie zamieszkałym przez około 1,3 miliona mieszkańców (Goodin, 2016). Stworzony kiedyś robak Stuxnet również został skierowany przeciwko fizycznemu celowi, jakim były wirówki przemysłowe, które zostały skutecznie uszkodzone (Chen i Abu-Nimeh, 2011). W licznych przeprowadzonych badaniach wykazano, że ogromna liczba przemysłowych systemów kontroli, nadzorujących operacje w bankach i szpitalach, jest podłączona bezpośrednio do internetu bez jakiejkolwiek ochrony (Kiravuo i in., 2015).

Broń nuklearna posiadana przez USA i Rosję prawdopodobnie już teraz mogłaby zabić większość ludzkości. Związek Radziecki prowadził również szeroko zakrojony program rozwoju broni biologicznej, z roczną zdolnością produkcyjną wynoszącą około 90–100 ton zmodyfikowanego wirusa prawdziwej ospy, a także genetycznie opracowanymi chorobami odpornymi na ciepło, zimno i antybiotyki (USAMRIID, 2014), które mogły spowodować ogromne liczby ofiar śmiertelnych w przypadku ich użycia. Rozwój inżynierii genetycznej i biologii syntetycznej umożliwił również tworzenie czynników biologicznych o wiele bardziej zabójczych od tych, które mogły kiedykolwiek ewoluować w sposób naturalny (ibid., s. 150–153). To, że jak dotąd żaden z tych scenariuszy się nie ziścił wynika z systemu wartości ludzi zajmujących kluczowe stanowiska, a nie dlatego, że powodowanie ogromnych szkód wymagałoby nadludzkich zdolności.

W dziedzinie manipulacji społecznych wykorzystano współczesne uczenie maszynowe do tworzenia prognoz opartych na „polubieniach" dawanych przez użytkowników na Facebooku, a prognozy te są dokładniejsze niż prognozy dokonywane przez znajomych na podstawie kwestionariusza osobowości (Youyou i in., 2015).

„Polubienia" zostały również wykorzystane do dokładnego przewidywania cech prywatnych, takich jak orientacja seksualna (Kosinski i in., 2013). Niektóre doniesienia w popularnej prasie podają, że firma marketingowa Cambridge Analytica wykorzystująca marketing oparty



Rys. 2. Różne ścieżki, w wyniku których SI może uzyskać DPS lub ZPS, prowadząc tym samym do katastrofy. Połączenia między węzłami oznaczają bramki LUB (pominięte w celu zwiększenia czytelności). Na przykład nadwyżka sprzętowa może wynikać albo z początkowego przełomu dotyczącego algorytmów LUB późniejszych optymalizacji. Jak omówiono w tekście, każda z nadwyżek sprzętowych, eksplozji prędkości i eksplozja inteligencji może przyczynić się do dwóch pozostałych, co zostało oznaczone ramką. Podobnie oznaczono różne kluczowe zdolności

na SI odegrała istotną rolę w wyborach prezydenckich w USA w 2016 roku oraz w referendum w sprawie członkostwa w Unii Europejskiej, które odbyło się w Wielkiej Brytanii w 2016 roku (Grassegger i Krogerus, 2017). Pomimo że prawdziwość tego twierdzenia pozostaje pytaniem otwartym i została zakwestionowana (Taggart, 2017), to daje to wyobrażenie, jakim rodzajem siły może dysponować SI zdolna do bardziej wyrafinowanego modelowania społecznego i manipulacji, umożliwiającego stworzenie świata, w którym o wynikach wyborów krajowych decydowałyby systemy SI.

Ogólnie rzecz biorąc, niektóre prawdopodobne możliwości, które mogą pomóc uzyskać MPS, to wojna biologiczna (rozwijanie i uwalnianie plag biologicznych), wojna cybernetyczna (atakowanie systemów kluczowej infrastruktury) i manipulacje społeczne (przekonanie wystarczająco wielu ludzi do wykonania woli SI, nawet tylko jeden człowiek może spowodować katastrofalne zniszczenia, jeśli byłby na przykład głową państwa). Należy zauważyć, że podobnie jak w przypadku inicjatorów wejścia w życie SI, posiadanie jednej zdolności może przyczynić się do posiadania innych. Na przykład SI zdolna do manipulacji społecznej może wykorzystać ją do znalezienia współpracowników zdolnych do działania w innych dziedzinach, a wojna cybernetyczna może dostarczyć kompromitujących informacji, które będą pomocne w szantażowaniu ludzi lub gromadzeniu informacji o ludzkim zachowaniu.

Zestawienie inicjatorów DPS/ZPS

Na rysunku 2 przedstawiono różne ścieżki, które mogą prowadzić do wcześniej omówionych katastrof. Każda z nich, eksplozja prędkości, eksplozja inteligencji lub nadwyżka sprzętowa, może przyczynić się do indywidualnego wejścia w życie, kiedy to pojedyncza SI osiągnie ogromne możliwości.

Nadwyżka sprzętowa może również przyczynić się do zbiorowego wejścia w życie SI, kiedy to dodatkowe możliwości sprzętowe mogą umożliwić tworzenie dużej liczby systemów SI w krótkim czasie, które następnie mogą zacząć ze sobą handlować, wkrótce wyprzedzając ludzkość. Węzeł „handlująca SI" to kolejny inicjator umożliwiający zbiorowe wejście w życie SI, reprezentujący podobny scenariusz, w którym jednak nie występuje nadwyżka sprzętowa, a różne kopie SI są budowane przez dłuższy okres, aż do momentu osiągnięcia poziomu zdolności niezbędnego do zbiorowego wejścia w życie. Każda forma wejścia w życie SI mogłaby doprowadzić do powstania SI z DPS. SI może również osiągnąć

DPS, jeśli ludzie dobrowolnie dadzą jej wystarczające możliwości.

Gdyby liczne SI otrzymały pewną władzę, niewystarczającą do osiągnięcia DPS, to nadal mogłyby osiągnąć ZPS. Ponadto nawet pojedyncza SI, która nie była wystarczająco silna do osiągnięcia DPS, mogłaby osiągnąć ZPS, gdyby posiadała pewne wystarczające zdolności ofensywne.

5. SI uzyskuje zdolność do samodzielnego działania

Ażeby SI stanowiła zagrożenie dla ludzkości, musi dysponować sposobami wpływania na świat i wywoływania katastrof. Powszechną propozycją ograniczenia potęgi SI jest próba ograniczenia jej zdolności do komunikowania się ze światem i wpływania na niego, co jest ogólnie znane jako „uwięzienie" lub „zapakowanie SI" (Chalmers, 2010, Armstrong i in., 2012, Yampolskiy, 2012, Bostrom, 2014).

Wyzwania związane z ograniczeniem SI są dwojakie. Po pierwsze, istnieje techniczne wyzwanie polegające na ograniczeniu SI w taki sposób, aby nie była w stanie się oswobodzić i nadal była w stanie dostarczać użytecznych informacji. Ponadto takie ograniczenie ma też wymiar społeczny, w którym decydenci mogą mieć różne zachęty do złagodzenia zabezpieczeń związanych z ograniczeniem SI lub nawet do całkowitego uwolnienia SI, nawet jeśli utrzymanie jej w zamknięciu byłoby technicznie wykonalne (Sotala i Yampolskiy, 2015). Jeśli uwięzienie ma być skuteczne, to muszą zostać spełnione wymagania zarówno techniczne, jak i społeczne.

Wyzwanie techniczne

Powszechną reakcją jest to, że wystarczająco inteligentna SI znajdzie pewien sposób na oswobodzenie się, albo przez socjotechnikę, albo przez znalezienie możliwych do wykorzystania słabości w zastosowanych fizycznych zabezpieczeniach. Możliwość ta została szeroko omówiona w wielu artykułach, w tym przez Chalmersa (2010) oraz Armstronga, Sandberga i Bostroma (2012). Ogólnie, autorzy są bardzo ostrożni w formułowaniu zdecydowanych twierdzeń na temat naszych zdolności do utrzymywania w ograniczeniu umysłu o wiele mądrzejszego niż nasz wbrew jego woli. Jednak przy ostrożnym projektowaniu nadal jest możliwe zaprojektowanie SI łączącej wewnętrzną motywację do pozostania w kontakcie z szeregiem zewnętrznych zabezpieczeń monitorujących SI.

Wyzwanie społeczne

Ograniczenie SI zakłada, że ludzie, którzy je tworzą i są za nie odpowiedzialni, muszą być faktycznie zmotywowani do ograniczenia SI. Jeśli grupa ostrożnych badaczy zbuduje i następnie z powodzeniem ograniczy stworzoną SI, może to nie odnieść zamierzonego skutku, jeśli inna grupa stworzy SI, która została celowo uwolniona od ograniczeń. Przyczyny pozbawienia ograniczeń SI mogą obejmować: (i) korzyści ekonomiczne lub presję konkurencyjną, (ii) przyczyny etyczne lub filozoficzne, (iii) zaufanie do zabezpieczeń SI oraz (iv) rozpaczliwe okoliczności, takie jak nieuchronna zagłada. Każdą z tych przyczyn omówiono poniżej.

Dobrowolne uwolnienie SI ze względu na korzyści ekonomiczne lub presję konkurencyjną

Jak wspomniano wcześniej w części „SI stopniowo przejmuje władzę", istnieje znaczna ekonomiczna zachęta do wdrażania systemów SI w celu kontroli korporacji. Może się to wydarzyć w dwóch formach: przez zwiększenie zakresu kontroli, jakim dysponują istniejące już systemy, albo alternatywnie przez aktualizację istniejących systemów lub dodawanie nowych z nieistniejącymi wcześniej możliwościami. Te dwie formy mogą się ze sobą łączyć. Jeśli pewne zadania wykonywane jak dotąd przez ludzi zostaną następnie przekazane ulepszonej SI, która stanie się zdolna do ich wykonywania, to może to zwiększyć autonomię SI zarówno przez zwiększenie jej zdolności, jak i zmniejszenie liczby ludzi biorących udział w dotychczasowym procesie.

Częściowym przykładem jest dążenie wojsk USA do ostatecznego przejścia do stanu, w którym ludzcy operatorzy broni robotycznej znajdowaliby się „nad pętlą", a nie „w pętli" (Wallach i Allen, 2013). Innymi słowy, podczas gdy dotychczas człowiek był zobowiązany do wyraźnego wydania polecenia, zanim robot mógł rozpocząć potencjalnie śmiercionośne działania, to w przyszłości ludzie mają po prostu nadzorować działania robota i interweniować w przypadku niekorzystnego rozwoju zdarzeń. Pozwoliłoby to systemowi na szybszą reakcję, jednak ograniczyłoby także możliwości ludzkich operatorów do podjęcia interwencji w przypadku błędów popełnianych przez system. Obecnie w przypadku licznych systemów wojskowych, takich jak automatyczne systemy obrony zaprojektowane do zestrzeliwania nadlatujących pocisków i rakiet, zakres ludzkiego nadzoru jest ograniczony do przyjęcia lub zastąpienia komputerowego planu działań w ciągu kilku sekund, co w praktyce może być za krótkim czasem na podjęcie sensownej decyzji (Human Rights Watch, 2012).

Sparrow (2016) przeanalizował trzy główne powody motywujące większe rządy do przejścia na autonomiczne systemy uzbrojenia i ograniczenie kontroli ludzi:

1. Obecnie istniejące, zdalnie pilotowane „drony wojskowe", takie jak US Predator i Reaper, wymagają dużej przepustowości łącza komunikacyjnego. Ogranicza to liczbę dronów, które mogą być rozmieszczone jednocześnie, i uzależnia je od satelitów komunikacyjnych, których nie ma każdy naród i które mogą zostać zablokowane lub zaatakowane przez wrogów. Konieczność stałej komunikacji ze zdalnymi operatorami uniemożliwia również tworzenie podwodnych dronów–okrętów, które musiałyby działać również w przypadku utraty łączności przed i podczas walki. Z tego powodu uczynienie dronów autonomicznymi i zdolnymi do działania bez nadzoru człowieka pozwoliłoby uniknąć tych wszystkich ograniczeń.
2. W szczególności w walce powietrznej zwycięstwo może zależeć od podjęcia bardzo szybkich decyzji. Już obecnie wymagania walki powietrznej znajdują się na granicy możliwości ludzkiego układu nerwowego, a dalszy postęp może zależeć od całkowitego usunięcia człowieka z tego procesu.

3. Większość rutynowych operacji dronów jest bardzo monotonna i nudna, co w znacznym stopniu przyczynia się do wypadków. Ponadto wydatki na szkolenia, wynagrodzenia i inne benefity dla operatorów dronów stanowią obecnie dużą część wydatków ponoszonych przez siły zbrojne.

Argumenty postawione przez Sparrowa są specyficzne dla dziedziny wojskowej, sugerują jednak, że „każda rozległa dziedzina dotycząca wysokich stawek, kontradyktoryjne podejmowanie decyzji oraz potrzeba szybkiego działania zostaną najprawdopodobniej coraz bardziej zdominowane przez systemy autonomiczne" (Sotala i Yampolskiy, 2015, s. 18). Podobne argumenty można wysunąć w dziedzinie biznesu. Wyeliminowanie ludzkich pracowników w celu zmniejszenia kosztów spowodowanych ich błędami i wynagrodzeniami mogłoby być kuszące dla firm. Już obecnie osiąganie zysków w dziedzinach transakcji o wysokiej częstotliwości zależy od osiągania lepszych wyników od innych traderów w ułamkach sekund. Pomimo że obecnie istniejące systemy SI nie są wystarczająco potężne, aby spowodować globalną katastrofę, to motywy, jak te przedstawione powyżej, mogą się przyczynić do ostatecznego podniesienia zdolności SI do takiego poziomu.

W przypadku braku wystarczających regulacji może dojść do „równania w dół ludzkiej kontroli", w którym podmioty państwowe lub biznesowe rywalizowałyby o ograniczenie kontroli ludzkiej i zwiększanie autonomii systemów SI w celu uzyskania przewagi nad konkurencją. Więcej szczegółów można znaleźć w pracy Armstronga i innych (2016), gdzie przedstawiono uproszczony scenariusz „wyścigu do przepaści". Byłoby to analogiczne do obecnej polityki „równania w dół", w której podmioty rządowe rywalizują o deregulację lub obniżenie podatków w celu utrzymania lub przyciągania przedsiębiorstw.

Ograniczenie systemów SI może także być argumentowane tym, że przyznanie systemom SI większych możliwości i autonomii może stwarzać znaczne ryzyko w przypadku nieprawidłowego działania SI. W biznesie ogranicza to zakres, w jakim duże i ugruntowane firmy mogą zaadoptować systemy kontroli opartej na SI, z drugiej strony startupy są zachęcane do inwestowania w autonomiczną SI, tak aby uzyskać przewagę nad konkurencją. W dziedzinie handlu algorytmicznego systemy SI mogą obecnie obracać ogromnymi sumami pieniędzy pomimo możliwości spowodowania znacznych strat. W 2012 roku Knight Capital straciła 440 mln USD z powodu usterki w oprogramowaniu transakcyjnym (Popper, 2012, Securities and Exchange Commission, 2013). Sugeruje to, że jeśli nawet nieprawidłowo działająca SI może potencjalnie powodować poważne ryzyko, to niektóre firmy nadal będą skłonne inwestować w powierzenie kontroli nad swoją działalnością autonomicznej SI, jeśli potencjalny zysk może być wystarczająco duży.

Prawo Stanów Zjednoczonych dopuszcza już możliwość nadania SI osobowości prawnej, ustawiając SI jako zarządzającą spółką z ograniczoną odpowiedzialnością. Człowiek może zarejestrować spółkę z ograniczoną odpowiedzialnością (z o.o.), zawrzeć umowę operacyjną określającą, że spółka z o.o. będzie zarządzana przez SI i następnie wystąpić z tej spółki (Bayern, 2015). Rezultatem tego jest podmiot prawny działający niezależnie i bez nadzoru oraz kontroli ze strony człowieka. Firmy kontrolowane przez SI mogą być również tworzone w różnych miejscach znajdujących się poza jurysdykcją USA. Ograniczenia zabraniające korporacjom aby nie miały właścicieli można w dużej mierze obejść, stosując takie sztuczki, jak tworzenie sieci korporacji, które są wzajemnymi właścicielami samych siebie (LoPucki, 2017). Możliwą początkową strategią mogłoby być opracowanie licznych systemów SI, wyposażenie ich w początkowe zasoby, a następnie uruchomienie kontroli nad własnymi korporacjami. W takim przypadku ryzykiem objęte są jedynie te początkowe zasoby z jednoczesną wizją potencjalnych zysków, jakie korporacja może uzyskać w przypadku odniesienia sukcesu. W przypadku odniesienia sukcesu przez takie korporacje i związanego z tym skutecznego osłabienia bardziej znanych firm, zostałaby wywarta presja na te firmy, aby one także przekazały kontrolę autonomicznym systemom SI.

Dobrowolne uwolnienie w celu osiągnięcia korzyści kryminalnych lub terroryzmu

LoPucki (2017) twierdzi, że jeśli człowiek stworzy autonomicznego agenta mającego ogólny cel, taki jak „optymalizacja zysku", a następnie agent ten niezależnie zdecyduje, by na przykład popełnić przestępstwo w celu zwiększenia zysku, to prokuratorzy mogą nie być w stanie skazać człowieka za to przestępstwo i jedynym zarzutem wobec człowieka może co najwyżej być oskarżenie o lekkomyślność. LoPucki utrzymuje, że ta „luka w odpowiedzialności" zapewnia między innymi, że ludzie stworzą kiedyś korporacje kierowane przez SI.

Ponadto LoPucki (2017, s. 16) utrzymuje, że takie „podmioty algorytmiczne" można tworzyć anonimowo, a osoby posiadające osobowość prawną mogą przyznać im szereg praw, takich jak możliwość „kupowania i dzierżawy nieruchomości, zawarcia umowy z legalnymi firmami, otwierania kont bankowych, składania pozwów w celu wyegzekwowania swoich praw lub kupowania rzeczy na Amazonie i zamawiania ich wysyłki". Jeśli podmiot algorytmiczny zostałby stworzony w celu takim, jak finansowanie lub przeprowadzanie aktów terrorystycznych, byłby wolny od presji społecznej lub zagrożeń ze strony ludzkich kontrolerów:

> Decydując się na próbę zamachu stanu, zbombardowanie restauracji lub zgromadzenie zbrojnej grupy w celu zaatakowania centrum handlowego, kontrolowana przez człowieka istota naraża życie swoich kontrolerów. Takie same decyzje podjęte przez podmiot algorytmiczny stwarzają ryzyko jedynie wobec zasobów, które podmiot algorytmiczny wydaje na planowanie i realizację (LoPucki, 2017, s. 18).

Podczas gdy większość grup terrorystycznych powstrzymałaby się przed celowym zniszczeniem świata, ograniczając się co najwyżej do spowodowania katastrofalnego ryzyka, to nie wszystkie z grup terrorystycznych mogłyby tak postąpić. Niektóre grupy mogą być zainteresowane spowodowaniem wyginięcia człowieka, w szczególności ekoterroryści uważający ludzkość za szkodliwą dla

planety oraz terroryści religijni uważający, że świat musi zostać zniszczony, aby osiągnąć zbawienie (Torres, 2016, 2017, rozdz. 4).

Dobrowolne uwolnienie ze względów estetycznych, etycznych lub filozoficznych

Kilku myślicieli, takich jak Gunkel (2012), poruszyło kwestię praw moralnych maszyn oraz tego, że nie wszyscy stanowczo uznają ograniczenie SI za etycznie dopuszczalne. Projektant wyrafinowanej SI może postrzegać ją jako coś w rodzaju swojego dziecka i czuć, że zasługuje ono na prawo do autonomicznego działania w społeczeństwie, bez jakichkolwiek zewnętrznych ograniczeń.

Dobrowolne uwolnienie z powodu zaufania zabezpieczeniom SI

Jeśli zespół badawczy ma ograniczyć SI, to musi poważnie potraktować możliwość, że stanie się ona niebezpieczna. Obecne badania nad SI nie obejmują żadnych zabezpieczeń ograniczających, ponieważ naukowcy mają uzasadnione przekonanie, że ich systemy nie są nawet zbliżone do ogólnej inteligencji. Wiele z tworzonych systemów jest również podłączonych bezpośrednio do internetu. Mamy nadzieję, że zabezpieczenia zaczną być wdrażane, gdy naukowcy stwierdzą, że tworzony system może mieć bardziej ogólne możliwości, będzie to jednak zależeć od ogólnej kultury bezpieczeństwa społeczności badawczej zajmującej się rozwojem SI (Baum, 2016), a w szczególności od konkretnej grupy badawczej. Jeśli grupa badawcza błędnie uzna, że jej SI nie może osiągnąć niebezpiecznego poziomu zdolności, to może nie zastosować wystarczających zabezpieczeń ograniczających.

Oprócz przekonania, że SI jest niewystarczająco zdolna do bycia zagrożeniem, badacze mogą również (poprawnie lub niepoprawnie) wierzyć, że udało się im dostosować SI do ludzkich wartości, tak aby nie miała żadnej motywacji do wyrządzania szkód ludziom.

Dobrowolne uwolnienie z desperacji

Miller (2012) zwraca uwagę, że jeśli ktoś byłby bliski śmierci, to z przyczyn naturalnych, będąc przegranym w wojnie lub z jakiegokolwiek innego powodu, mógłby uwolnić nawet potencjalnie niebezpieczny system OSI. Byłby to racjonalny sposób działania, o ile ten ktoś ceniłby sobie przede wszystkim własne przetrwanie i sądził, że nawet niewielka szansa na uratowanie życia przez OSI była lepsza niż niemal pewna śmierć.

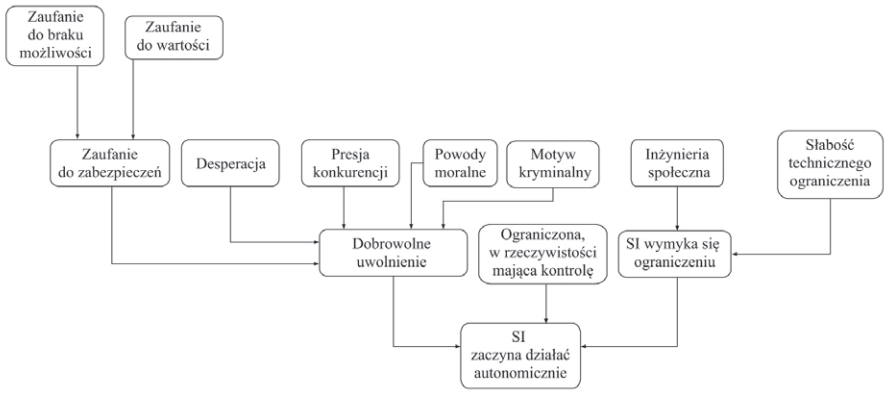
SI pozostaje ograniczona, jednak ostatecznie przejmuje kontrolę

Nawet jeśli ludzie technicznie występowaliby w pętli procesu decyzyjnego, to mogliby nie mieć czasu, okazji, motywacji, inteligencji lub pewności siebie, aby zweryfikować porady udzielone przez SI. Byłoby tak w szczególności w przypadku, gdyby SI działała już przez pewien czas i zyskała reputację godnej zaufania. Automatyczną reakcją na zalecenia SI może stać się rutynowa praktyka i coraz trudniejsze może być zakwestionowanie „autorytetu" jej zaleceń. W rezultacie SI mogłaby efektywnie narzucać własne decyzje (Friedman i Kahn, 1992).

Podobnie Bostrom i Yudkowsky (2014) zwracają uwagę, że współcześni biurokraci bardzo często dosłownie przestrzegają ustalonych procedur, zamiast dokonywać własnych osądów, w obawie o to, że mogliby zostać później obwinieni za popełnione błędy. Podobnym sposobem na unikanie winy mogłoby być posłuszne przestrzeganie wszystkich zaleceń systemu SI.

O'Neil (2016) udokumentował wiele sytuacji, w których współczesne uczenie maszynowe jest wykorzystywane do podejmowania merytorycznych decyzji, nawet jeśli dokładne modele stojące za tymi decyzjami mogą być tajemnicą handlową lub w inny sposób ukryte przed krytyką zewnętrzną. Między innymi takie modele były już wykorzystywane do zwalniania nauczycieli sklasyfikowanych przez system jako nieefektywni oraz do wymierzania surowszych wyroków przestępcom, których model określił jako obarczonych wysokim ryzykiem ponownego popełnienia przestępstwa. W niektórych przypadkach ludzie byli sceptycznie nastawieni do wyników takich systemów i nawet wskazywali prawdopodobne powody błędności wyników, nadal jednak zgadzali się z autorytetem systemu, o ile nie można było jednoznacznie wykazać, że model się pomylił.

W dziedzinie wojskowej Wallach i Allen (2013) zasygnalizowali istnienie robotów, które próbują automatycznie wykrywać lokalizacje wrogich snajperów i wskazywać je żołnierzom. W zakresie, w jakim ci żołnierze zaczęli ufać tym robotom, można postrzegać ich jako wykonujących rozkazy robotów. W końcu wyposażenie robota we własną broń po prostu wyeliminowałoby formalną potrzebę, by to człowiek pociągał za spust. Na rysunku 3 przedstawiono podsumowanie różnych sposobów, w jakie SI może uzyskać swobodę autonomicznego działania.



Rys. 3. Sposoby, w wyniku których SI może uzyskać swobodę autonomicznego działania. Połączenia pomiędzy węzłami oznaczają bramki LUB (pominięte w celu zwiększenia czytelności): na przykład zaufanie do zabezpieczeń może wynikać z zaufania do braku możliwości LUB zaufania do wartości

6. Uwagi na temat pojedynczej i licznej SI

Wiele analiz koncentruje się na przypadku istnienia tylko pojedynczej SI. Scenariusz, w którym istotna byłaby tylko jedna kopia SI, mógłby się wydarzyć, gdyby:

1. Pierwsza stworzona SI bardzo szybko osiągnęłaby DPS, zaraz po jej utworzeniu.
2. Pewna grupa badawcza znacznie wyprzedziła wszystkich konkurentów w rozwoju SI i była w stanie utrzymać tę przewagę przez dłuższy czas.

Na potrzeby tej analizy przyjęto scenariusz, w którym istnieje wiele kopii pojedynczej SI, wszystkie z nich mają te same cele, a cała ich zbiorowość jest traktowana jako pojedyncza SI. To samo dotyczy sytuacji, w której pojedyncza SI tworzy bardziej wyspecjalizowane „robotnicze SI", aby zrealizować jakiś bardziej określony cel związany z osiągnięciem celu podstawowego.

Spośród dwóch powyższych możliwości opcja druga wydaje się stosunkowo mało prawdopodobna w ciągu co najwyżej kilku lat, biorąc pod uwagę obecną silną konkurencję w dziedzinie rozwoju SI. Pomimo że jedna firma mogłaby osiągnąć znaczącą przewagę w pewnej rzadkiej niszy przy niewielkiej konkurencji, to wydaje się, że nie zdarzy się to w przypadku rozwoju SI.

Możliwym wyjątkiem może być sytuacja, gdy firmie uda się całkowicie zmonopolizować pewną dziedzinę lub jeśli będzie miała zasoby programistyczne, jakich nie ma nikt inny. Na przykład firmy, takie jak Google i Facebook, mają obecnie dostęp do znacznie większych zbiorów danych niż większość innych podmiotów korporacyjnych lub akademickich. We współczesnym uczeniu maszynowym duże zestawy danych w połączeniu z prostymi modelami zwykle dają lepsze wyniki niż małe zestawy danych i bardziej wyrafinowane modele (Halevy i in., 2009). Jak zauważyli Goodfellow i inni (2016, rozdz. 1), algorytm głębokiego uczenia wymaga z reguły co najmniej 10 milionów oznakowanych przykładów w celu osiągnięcia wydajności na poziomie człowieka lub lepszej.

Z drugiej strony zależność od tak ogromnych zestawów danych jest dziwactwem obecnych technik uczenia maszynowego. Ludzie uczą się na podstawie znacznie mniejszych ilości danych, a także są w stanie wykorzystywać swój proces uczenia się w bardziej elastyczny sposób, co sugeruje fundamentalne różnice w sposobie, w jaki ludzie i współczesne algorytmy uczą się (Lake i in., 2016). Z tego powodu możliwe jest, że OSI byłaby w stanie uczyć się na podstawie znacznie mniejszych ilości danych, a projekt OSI nie byłby tak ograniczony przez potrzebę dużych zbiorów danych[10].

Innym prawdopodobnym kluczowym zasobem mogą być zasoby sprzętowe. Być może pierwsza OSI będzie wymagała ogromnych mocy obliczeniowych. Bostrom (2017) zauważa, że jeśli w rozwoju SI istnieje duży stopień otwartości i każdy ma dostęp do tych samych algorytmów, to właśnie sprzęt może się stać głównym czynnikiem ograniczającym. Gdyby wymagania sprzętowe dla SI były stosunkowo niskie, wysoka otwartość mogłaby doprowadzić do powstania wielu jednostek SI. Z drugiej strony, jeśli sprzęt byłby głównym czynnikiem ograniczającym i potrzebne byłyby duże ilości sprzętu, to kilka zamożnych organizacji mogłoby przez jakiś czas zmonopolizować SI. Jak wcześniej omówiono w części „Inicjatorzy katastroficznych zdolności", optymalizacje oprogramowania mogą szybko zmniejszyć zapotrzebowanie na sprzęt, ograniczając tym samym czas, kiedy sprzęt może być kluczowym ograniczeniem.

Branwen (2012) zasugerował, że produkcja sprzętu zależy od niewielkiej liczby scentralizowanych fabryk, które byłyby łatwym celem regulacji. Sugerowałoby to możliwą drogę, według której SI mogłaby podlegać regulacjom rządowym, ograniczając liczbę wdrożonych jednostek SI. Podobnie pojawiły się propozycje rządowych i międzynarodowych regulacji rozwoju SI (np. Wilson, 2013, argumentów przeciwko szukaj w: McGinnis, 2010). W przypadku pomyślnego uchwalenia, takie regulacje mogą ograniczyć liczbę wdrożonych jednostek SI.

Innym możliwym kluczowym zasobem byłoby posiadanie nieoczywistego przełomowego osiągnięcia, które byłoby trudne do odkrycia dla innych badaczy. Gdyby było ono utrzymywane w tajemnicy, to jedna firma mogłaby prawdopodobnie znacznie posunąć się naprzód w stosunku do innych.

Skuteczne procedury ograniczania SI mogą również zwiększać szanse na powstanie wielu SI, ponieważ ograniczenie pierwszych jednostek SI, umożliwiłoby innym projektom nadrobienie zaległości.

Sytuacja rozwoju wielu różnych jednostek SI może zaistnieć, gdy:

1. Kilku twórców osiągnęło zdolność do budowania SI w tym samym czasie i żadna SI nie osiągnęła DPS.
2. Jeden twórca mógł wyprodukować kilka różnych SI mających różne cele.
3. Tylko jeden twórca był w stanie wdrożyć SI, ale ta SI stworzyła własne kopie i nie dostosowała celów tych kopii do własnych.

Trudno przewidzieć konsekwencje istnienia wielu jednostek SI. Obecnie opracowywana jest SI w celu ostrzegania przed potencjalnym ryzykiem, na przykład przez przewidywanie ryzyka finansowego na podstawie artykułów prasowych (Rönnqvist i Sarlin, 2017), a od wielu lat wykorzystuje się SI do celów takich jak automatyczne wykrywanie włamań (Lunt, 1988). Bardziej wyrafinowana i dopasowana do człowieka SI może pomóc w obronie przed niedopasowanymi systemami SI (Hall, 2007, Goertzel i Pitt, 2012).

Z drugiej strony podstawowym problemem związanym z obroną jest to, że aby zapobiec katastrofie, obrońcy muszą odnieść sukces za każdym razem, podczas gdy atakującemu wystarczy tylko jedno odniesienie sukcesu. W przypadku istnienia licznych SI procedury, takie jak ograniczanie SI, musiałyby być skuteczne dla każdej pojedynczej SI, a wszyscy ludzie musieliby uznawać stosowanie ograniczeń SI za wartościowe. W rezultacie istnienie licznych SI jest zwielokrotnieniem liczby systemów, które mogłyby potencjalnie spowodować katastrofę.

Inną kwestią jest to, że istnienie licznych SI wydaje się pomocne tylko wtedy, gdy wystarczająco duża ich część ma wartości dostosowane do wartości ludzkich. Scenariusz z istniejącą liczną SI, z których każda realizuje interesy w niewielkim stopniu związane z wartościami ludzkimi, najprawdopodobniej byłby niekorzystny dla ludzkich wartości. Zwłaszcza jeśli wszystkie SI byłyby znacznie bardziej zdolne niż ludzie, to taki scenariusz po prostu stawia ludzi w krzyżowym ogniu.

7. Wnioski

W tym rozdziale rozważaliśmy różne drogi rozwoju SI, które mogą zakończyć się katastrofą (tabela 2). W części „Inicjatorzy katastrofy" przedstawiono dowody na to, że nadmierne skupianie się na SI osiągającej DPS umożliwiające jej osiągnięcie całkowitej dominacji nad światem, może być nierozsądne. Wydaje się raczej uzasadnione, aby rozważyć również możliwości uzyskania ZPS, poziomu zdolności, który może umożliwić SI spowodowanie co najmniej dziesiątek milionów ofiar. Oprócz tego jest znacznie bardziej prawdopodobne, że SI uzyska ZPS niż DSA, a chaos spowodowany przez SI z ZPS może ostatecznie doprowadzić do pojawienia się SI z DSA, nawet jeśli pierwsza SI zostałaby pomyślnie wyłączona.

Tabela 2. Różne drogi prowadzące do katastroficznych scenariuszy

Poziom strategicznej przewagi SI	• Decydujący • Znaczny
Próg zdolności SI do wystąpienia braku współpracy	• Bardzo niski do bardzo wysokiego, w zależności od różnych czynników
Źródła zdolności SI	• Indywidualne wejście w życie • Nadwyżka sprzętowa • Eksplozja prędkości • Eksplozja inteligencji • Zbiorowe wejście w życie • Kluczowe możliwości • Wojna biologiczna • Wojna cybernetyczna • Manipulacja społeczna • Coś innego • Stopniowe przesunięcie władzy i możliwości
Sposoby SI na osiągnięcie autonomii	• Oswobodzenie się • Manipulacja społeczna • Słabość techniczna • Dobrowolne uwolnienie • Przyczyny ekonomiczne lub konkurencyjne • Przyczyny kryminalne lub terrorystyczne • Przyczyny etyczne lub filozoficzne • Desperacja • Zbytnie zaufanie • § Do braku możliwości • § Do wartości • Ograniczona, w rzeczywistości mająca kontrolę
Liczba SI	• Pojedyncza • Wiele

Rozważenie scenariuszy, w których SI osiąga „tylko" ZPS wymaga położenia większego nacisku na analizę, kiedy SI byłaby skłonna zaryzykować podjęcie działań wrogich wobec ludzi. Liczne rozważania przedstawiono w części „Kiedy zostaną podjęte działania przeciwko przewadze strategicznej". Zasadniczo, jeśli SI działałaby racjonalnie, to zainicjowałaby agresywne działania tylko wtedy, gdyby spodziewana uzyskana w ten sposób użyteczność przewyższała spodziewaną użyteczność uzyskaną w przypadku współpracy, przy uwzględnieniu ryzyka niepowodzenia i odpowiadającego mu odwetu ze strony ludzi (Shulman, 2010). Istnieje jednak wiele sytuacji, które mogą zmusić SI do podjęcia wrogiego działania.

Próbując ustalić katastrofalne ryzyko związane z SI jako formę ryzyka rozłącznego, gdzie wiele różnych spraw może potoczyć się niekorzystnie, w części „Inicjatorzy katastroficznych zdolności" przedstawiono różne sposoby, dzięki którym SI lub grupy SI mogą się stać wystarczająco zdolne do uzyskania pewnej formy przewagi strategicznej. Omówiono indywidualne scenariusze wejścia w życie wraz z trzema głównymi podtypami, scenariusze zbiorowego wejścia w życie, scenariusze, w których władza jest przejmowana przez systemy SI, oraz scenariusze, w których SI staje się wystarczająco zdolna, by zdobyć kluczowe możliwości dające jej ZPS lub DPS.

Ponieważ SI może stać się zdolna tylko wtedy, gdy uzyska wystarczającą autonomię, w części „SI uzyskuje zdolność do samodzielnego działania" przedstawiono różne sposoby, w jakie

SI może osiągnąć taką autonomię. Przedstawione przyczyny przyznania autonomii SI obejmowały: (i) korzyści ekonomiczne lub presję konkurencyjną, (ii) przyczyny kryminalne lub terrorystyczne, (iii) przyczyny etyczne lub filozoficzne, (iv) zaufanie do zabezpieczeń SI oraz (v) rozpaczliwe okoliczności, takie jak wizja nieuchronnej zagłady. Ponadto wystarczająco inteligentna SI może uniknąć ograniczenia lub może stać się wystarczająco wpływowa, aby uzyskać skuteczną kontrolę nawet pomimo teoretycznego istniejącego ograniczenia.

Wreszcie, wszystkie drogi prowadzące do katastrofy mogą ulegać zwielokrotnieniu w przypadku istnienia licznych różnych kopii SI, z których każda może być w stanie osiągnąć autonomię, a następnie duży poziom zdolności. W części „Uwagi na temat pojedynczej i licznej SI" omówiono, czy możemy się spodziewać bardzo małej liczby SI, czy też będzie ich wiele, a także niektóre implikacje w stosunku do każdego scenariusza.

Łączenie różnych dróg omówionych w poprzedniej części może skutkować wieloma różnymi scenariuszami (patrz ramka poniżej), poczynając od tych, w których SI oswobadza się i szybko osiąga superinteligencję, po te, w których SI jest budowana celowo z zamiarem kontrolowania korporacji, a rosnące zasoby są jej dobrowolnie przydzielane aż do momentu, gdy SI zawładnie całą planetą. Każda z tych dróg będzie musiała zostać osobno oceniona pod kątem wiarygodności, a także pod kątem najbardziej odpowiednich metod zapobiegających. Mamy nadzieję, że taka analiza pozwoli wykorzystać pozytywny potencjał SI, jednocześnie unikając katastrofy.

8. Niektóre przykładowe scenariusze

Różnorodne kombinacje różnorakich omówionych ścieżek mogą prowadzić do powstania wielu rodzajów scenariuszy ryzyka związanego z rozwojem SI. Poniżej przedstawiono cztery przykłady:

- **Klasyczne przejęcie**
  (Decydująca przewaga strategiczna, wysoki próg zdolności, eksplozja inteligencji, wejście w życie SI i pojedyncza SI)

  „Klasyczny" scenariusz przejęcia SI został opisany przez Bostroma (2014, rozdz. 6). Rozwijana SI ostatecznie staje się lepsza w projektowaniu SI niż jej programiści. SI wykorzystuje tę zdolność do eksplozji inteligencji i ostatecznie ucieka do internetu ze swojego ograniczonego środowiska. Po sekretnym zdobyciu wystarczającego wpływu i zasobów przeprowadza atak przeciwko ludzkości, eliminując ludzkość jako dominującego gracza na Ziemi, w wyniku czego SI może bez przeszkód realizować własne plany.

- **Stopniowe przejęcie**
  (Zasadnicza przewaga strategiczna, wysoki próg zdolności, stopniowe przesunięcie władzy, uwolnienie z przyczyn ekonomicznych i wiele kopii SI)

  Wiele korporacji, rządów i osób prywatnych dobrowolnie powierza wykonanie zadań SI, aż do momentu zupełnego uzależnienia od systemów AI. W początkowym etapie są to wyspecjalizowane systemy SI, jednak ciągłe aktualizacje sprawiają, że niektóre z nich osiągają poziom ogólnej inteligencji. Stopniowo zaczynają one podejmować wszystkie decyzje. Wiemy, że pozwolenie im na prowadzenie takich działań jest ryzykowne, jednak są one zaangażowane w zbyt wiele spraw, które przynoszą zysk i są naprawdę skuteczne w tworzeniu pożytecznych dla ludzkości przedmiotów. Do pewnego czasu.

- **Wojny zdesperowanych SI**
  (Zasadnicza przewaga strategiczna, niski próg zdolności, kluczowe zdolności, oswobodzenie się SI i wiele kopii SI)

  Wielu różnych twórców opracowuje systemy SI. Większość tych prototypów nie jest zgodna z ludzkimi wartościami i nie posiada niezwykłych zdolności, jednak liczne z tych SI uważają, że niektóre inne prototypy mogą okazać się bardziej zdolne. W rezultacie systemy SI starają się zdradzić ludzkość nawet pomimo małych szans na powodzenie, motywowane tym, że miałyby jeszcze mniejsze szanse na osiągnięcie swoich celów, gdyby nie zdradziły. Społeczeństwo zostaje zaatakowane przez różne systemy wymykające się spod kontroli, które mają kluczowe możliwości do wyrządzenia katastrofalnych szkód, zanim zostaną powstrzymane.

- **Czy ludzkość uważa, że ma szczęście?**
  (Decydująca przewaga strategiczna, wysoki próg zdolności, kluczowe zdolności, ograniczona jednak w efekcie mająca kontrolę SI i pojedyncza SI)

  Google zaczyna podejmować decyzje dotyczące wprowadzanych produktów i strategii zgodnie z wytycznymi strategicznej SI. Pozwala to firmie stać się jeszcze potężniejszą i bardziej wpływową, niż jest obecnie. Kierując się strategią, SI zaczyna podejmować coraz bardziej wątpliwe działania, które zwiększają jej władzę i możliwości. W końcu staje się zbyt potężna, aby społeczeństwo mogło ją powstrzymać. Trudny do zrozumienia kod napisany przez strategię SI wykrywa i subtelnie sabotuje projekty SI innych twórców, aż do momentu, kiedy Google nie stanie się dominującą potęgą światową. Odmiana tego scenariusza z gwałtownym wejściem w życie SI została opisana w rozdziale otwierającym pracę Tegmarka (2017).

Przypisy

1. Na przykład Goertzel (2015) krytykuje Bostroma (2014): „To, co znajdujemy w Superintelligence jest ostrożnym filozoficznym sformułowaniem argumentującym, dlaczego katastrofalne rezultaty są możliwe, a następnie bardziej praktycznym przewidywaniem na podstawie „najgorszego planu", odrzucając pozytywne możliwości".
2. Yampolskiy (2015) przedstawił również taksonomię tego, w jaki sposób SI może mieć wartości, które nie są dostosowane do ludzkich, jednak jest to jedynie ogólna taksonomia, a nie bardziej szczegółowa analiza przyczyn.
3. Określenia „motywacja" używa się tutaj w ogólnym znaczeniu i nie należy go traktować jako twierdzenia, że SI miałaby system motywacyjny podobny do ludzkiego. Zamiast przyjmować założenia

mechanizmów leżących u podstaw Si, zakładamy, że jej zachowanie można z łatwością przewidzieć, zakładając postawę celową (Dennett, 1971, 2009), w której przyjmuje się, że zachowanie systemu można wytłumaczyć celami i przekonaniami. Na przykład, pomimo że obliczenia komputera do gry w szachy praktycznie nie mają nic wspólnego z ludzkim myśleniem, to jego ruchy można nadal skutecznie przewidzieć, zakładając, że „chce" wygrać w szachy i „zna" zasady szachowe. Daje to podstawę do przewidywania, że zawsze wybierze z listy wykonalnych ruchów taki, który najbardziej zbliża do celu wygrania gry. Pomimo że najlepszy ruch może nie być oczywisty, to przyjęcie zamierzonej postawy nadal pozwala ludzkiemu obserwatorowi poprawić przewidywania dotyczące tego, co zrobi komputer, eliminując oczywiste złe ruchy (ibid.).

4. Może się to zdarzyć na przykład wtedy, gdy ludzie nie są pewni, czy cele SI rzeczywiście uwzględniają wszystko co ma wartość dla ludzi.
5. SI, która po prostu dążyłaby do maksymalizacji ludzkiego szczęścia w ogóle, mogłaby być gotowa poświęcić wszystkich obecnie żyjących ludzi, gdyby pomyślała, że pozwoli to na stworzenie bardziej szczęśliwych ludzi później.
6. Przykładem wartości, która nie była samozachowawcza jako cel instrumentalny, jest sztuczna inteligencja chcąca (z jakiegokolwiek powodu) sama się zniszczyć.
7. Shulman (2010) zaproponował projekt, który osiąga niemal maksymalną użyteczność po otrzymaniu stałego sygnału nagrody od ludzi. Zrobienie czegokolwiek, co zagrażałoby ludziom, wyłączałoby ten sygnał nagrody, co wiązałoby się z ryzykiem utraty większości użyteczności SI i tym samym zniechęcałoby ją do podejmowania wrogich działań. Bostrom (2014) omówił możliwość powiązania funkcji nagrody SI ze strumieniem wstępnie wygenerowanych tokenów kryptograficznych, które można łatwo zniszczyć w przypadku podjęcia wrogiej akcji przez SI. Podjęcie wrogiej akcji byłoby wtedy realną strategią tylko wtedy, gdyby SI była bardzo pewna, że będzie w stanie przejąć zapasy tokenów, zanim zostanie zniszczona. Wysoka stopa zwrotu, która sprawiałaby, że SI traktowałaby priorytetowo nagrody krótkoterminowe nad długoterminowymi mogłaby również uniemożliwić podejmowanie działań, które bezpośrednio nie przyczyniałyby się do uzyskania nagród (Shulman, 2010). SI, która miałaby inną formę „trywialnego" lub łatwego do spełnienia celu, lub której wyraźnym celem byłby niewielki wpływ i niewywieranie znaczącego wpływu na świat (Armstrong i Levinstein, 2017), również byłaby bardziej skłonna współpracować i unikać kontragresji. Wszystkie te propozycje są jednak obecnie spekulacyjne i nie jest jasne, jak dobrze działałyby w przyszłości.
8. Wymagany do tego stopień inteligencji jest niejasny. Przejście ze scentralizowanej SI do rozproszonego systemu składającego się ze skoordynowanych subagentów może wymagać zaawansowanych umiejętności projektowych, jednak nie wymaga tego po prostu skopiowanie oryginalnej SI. Takie kopie mogą nie być optymalnie skoordynowane ze sobą, ale jeśli nie miałyby interesu własnego i były skupione na wspólnym celu, to mogą nadal współpracować bardziej skutecznie niż grupy ludzi, których współpracę utrudniają jednostki (Olson, 1965) i grupy dbające o własny interes (DeScioli i Kurzban, 2013, Greene, 2013). Ponadto SI od samego początku mogłaby również zostać zaprojektowana jako system rozproszony.
9. Jednak to, czy można nakreślić znaczącą różnicę pomiędzy „indywidualną SI" i „społecznością SI", jest nieco niejasne. Systemy SI mogą nie mieć indywidualności w takim samym sensie jak ludzie, szczególnie jeśli mają wysoką przepustowość komunikacji w stosunku do zdolności obliczeniowej wewnątrz węzła.
10. Z drugiej strony istnieją teorie, które sugerują, że ludzka zdolność do szybkiego uczenia się może wynikać z układów neuronów kodujących z dużą ilością odziedziczonych, wcześniej istniejących informacji. Opracowanie podobnych startowych „danych ładowania początkowego" dla SI może ponownie wymagać dużych zbiorów danych. Na przykład H. Barrett i Kurzban (2006) zauważają, że takie wrodzone systemy zostały zaproponowane do wykrywania oszustów, języka, teorii umysłu, orientacji przestrzennej, liczby, mechaniki intuicyjnej, emocji, wykrywania krewnych i rozpoznawania twarzy, a Spelke i Kinzler (2007) twierdzą, że ludzkie poznanie zbudowane jest na czterech podstawowych systemach wiedzy reprezentowania obiektów, działań, liczb i przestrzeni.

Bibliografia dostępna pod linkiem:
nis.com.pl/bibliografia.html

Fragment pochodzi z książki:
Sztuczna inteligencja. Bezpieczeństwo i zabezpieczenia,
Roman V. Yampolskiy (redakcja).
Wydawnictwo Naukowe PWN,
Warszawa 2020