ANOTACIÓNS
sobre literatura e filosofía
nº 19, setembro de 2018

Miguel Penas

"O fracaso da Intelixencia Artificial computacionalista. A posibilidade dunha concepción corporal da cognición"

Euseino?

Anotacións 19

Anotacións sobre literatura e filosofía
nº 19, setembro de 2018

O fracaso da Intelixencia Artificial computacionalista. A posibilidade dunha concepción corporal da cognición

ANOTACIÓNS
sobre literatura e filosofía
nº 19, setembro de 2018

Miguel Penas

"O fracaso da Intelixencia Artificial computacionalista. A posibilidade dunha concepción corporal da cognición"

Euseino?

Primeira edición (PDF), setembro de 2018
ISSN 2340-8537
PUBLICACIÓN NON VENAL

Edición
Beatriz Fraga Cameán

Euseino? Editores
Fundación Euseino?
Rúa do Brasil 40-42, 5º Esda. 36204 Vigo, Galicia
http://euseino.org

A oposición que se erixiu entre a cultura e a técnica, entre o home e a máquina, é falsa e sen fundamento; non agocha máis que ignorancia ou resentimento. A cultura compórtase co obxecto técnico como o fai o ser humano co estranxeiro cando se deixa levar pola xenofobia primitiva.

Gilbert Simondon, 1958

En 1968, durante a presentación perante a prensa de *2001: A Space Odissey*, Marvin Minsky, asesor de Stanley Kubrick na película, proclamou: "Nunha xeración teremos computadoras intelixentes coma HAL na película *2001*" (Dreyfus 2, 331).

Xa pasou medio século dende aquela e a consumación do proxecto da Intelixencia Artificial (IA), que sería acadada segundo Minsky pola seguinte xeración de investigadores e segundo a insinuación un pouco máis pesimista de Kubrick en *2001*, non se produciu. O obxectivo do presente ensaio é ofrecer, dende unha perspectiva filosófica, unha explicación do fracaso desas aspiracións presentes na IA e amosar que as causas dese fracaso, que se poden resumir nunha certa concepción da intelixencia que guiou a IA durante décadas, tentan ser superadas por

un novo paradigma xurdido no seo das ciencias cognitivas. Este novo paradigma tenta levar a cabo o proxecto da IA dende uns presupostos radicalmente diferentes. Rexeitando a vella idea computacionalista de que é posible crear mentes sen corpo, presente aínda en boa parte dos achegamentos actuais á IA, asistiremos ao nacemento dunha nova concepción na que a mente, o corpo e a súa relación co ambiente aparecen como inseparables.

ACHEGAMENTO HISTÓRICO Á INTELIXENCIA ARTIFICIAL

Podemos definir a IA como a tentativa de simular o comportamento intelixente humano nalgún tipo de artefacto —xa sexa unha computadora dixital ou análoga, un robot ou os algoritmos evolucionarios propios das redes neuronais artificiais—. Na súa orixe posuía un propósito dobre que, a partir da década de 1970, supuxo unha bifurcación de camiños: por un lado tiña unha utilidade teorética, xa que constitúe unha ferramenta que permitiría comprender a natureza da intelixencia humana e probar empiricamente as teorías dispoñibles sobre ela; por outro lado, fai

posible a creación de ferramentas e sistemas cunha utilidade práctica (de feito, a principal fonte de financiación do Laboratorio de IA do Massachusetts Institute of Technology foi, na súa orixe, o Ministerio de Defensa estadounidense). O feito de que ambas finalidades se bifurcaran xa é un síntoma das dificultades ás que se enfronta a IA pois, malia non posuírmos unha explicación integral e satisfactoria do que é a intelixencia, a creación de utilidades prácticas en dominios concretos tivo un éxito relativo. Dado que o interese aquí é ofrecer unha comprensión filosófica da IA, centrarémonos na primeira finalidade.

O primeiro traballo dentro do campo da IA, aínda que entón non existía a disciplina con ese nome, é a creación en 1943 dun modelo de neuronas artificiais por parte de Warren McCulloch e Walter Pitts. No seu célebre artigo publicado en 1950, Alan Turing formulou de maneira explícita a posibilidade da IA ao preguntarse se as máquinas poden pensar. Para facelo, ideou a coñecida Proba de Turing, un sinxelo experimento que serviría para determinar se a actuación dunha máquina é indistinguible da dun ser humano. Nesta proba, manten-

se un diálogo, no que se pode mentir, entre un xuíz que está nun cuarto e unha persoa e unha máquina que están noutro. Se o xuíz é incapaz de distinguir quen é a persoa e quen é a máquina, entón esta última superaría a proba e podería considerarse intelixente.* O filósofo John Searle contestou a esta versión computacionalista da intelixencia cun contra-experimento denominado o cuarto chinés no que tenta amosar que unha máquina podería pasar a Proba de Turing sen ter ningunha comprensión do que fixo e sen que poida, por tanto, ser considerada intelixente.

Dende a década de 1950 ata mediados da de 1970, a investigación en IA estivo guiada polo paradigma triunfante nas ciencias cognitivas, o computacionalismo, baseado na idea de que o funcionamento da nosa mente é análogo ao das computadoras. O computacionalismo concibe a mente como un procesador de información que efectúa representacións simbólicas dos *inputs* recibidos por ela e que

* As probas CAPTCHA coas que estamos familiarizados hoxe en día e que tanto detestamos son unha versión desa Proba de Turing, tal como se recolle no seu acrónimo: *Completely Automated Public Turing test to tell Computers and Humans Apart* (Proba de Turing completamente pública e automatizada para diferenciar ordenadores e seres humanos).

manexa esas representacións por medio dunhas regras sintácticas, o que lle permite xerar os *outputs* desexados. A analoxía coas computadoras é clara, aínda que é necesario distinguir entre os dous tipos de computadoras existentes polo de agora: as análogas, cuxas unidades de información radican nun soporte físico e son por tanto continuas e as dixitais que, tal como indica o seu nome, formalizan a información en unidades numéricas que só posúen dúas posicións, un e cero, e son por tanto discretas —cada unidade mínima é unha entidade illada cunha posición fixa.

O modelo que triunfou historicamente é o das computadoras dixitais e iso, como imos ver, terá serias consecuencias para a IA pois vai implicar que a simulación da intelixencia sexa modelada en función dunhas determinadas posibilidades abertas pola enxeñería computacional. A intelixencia é concibida dunha maneira intelectualista como o manexo, por medio dunhas regras sintácticas, de símbolos que representan ou están-por aspectos illados e ben definidos da realidade que se pretende formalizar. O uso de computadoras dixitais en IA implica unha concepción da realidade como algo que se pode descompoñer ou analizar en unidades illadas e indepen-

dentes (obxectos e propiedades) que son expresables nunha linguaxe formal. A cognición é concibida como un proceso abstracto, simbólico, representacional, cuxas unidades de información son independentes do contexto: sóñase cunha linguaxe formal neutral, libre de toda ambigüidade e por extensión de toda interpretación, que poida representar a realidade. Sóñase coa posibilidade da redución do comportamento intelixente humano, isto é, o seu saber-facer (*know-how*) a un simple saber-que (*knowwhat*), un conxunto de feitos e regras que se poida aplicar en todos os casos.

Sen entrar nunha exposición detallada do desenvolvemento histórico da IA, imos facer un percorrido que nos permita subliñar os principais problemas presentes nas súas aspiracións. Dende a década de 1950 ata a de 1970, as investigacións, guiadas polo computacionalismo, dirixíronse principalmente aos seguintes ámbitos: a resolución de problemas (*General Problem Solving,* de Newell e Simon), o recoñecemento de formas (*pattern recognition*), a tradución de linguas (por exemplo, o programa STUDENT de Bobrow), a creación de micro-mundos nos que os programas se poidan desenvolver (o que máis éxito tivo foi o SHRDLU

de Winograd, capaz de entender ordes e de responder adecuadamente a elas en inglés dentro dun micromundo de bloques) e a creación de programas para xogar a determinados xogos, especialmente o xadrez.

En xeral, a metodoloxía seguida foi a de tentar avanzar en dominios restrinxidos coa esperanza de que as técnicas acadadas se puidesen aplicar en ámbitos cada vez máis xerais. Efectivamente, acadáronse uns éxitos iniciais pero o funcionamento das computadoras presenta problemas á hora de tentar xeneralizar eses éxitos. A aplicación do coñecemento a un determinado ámbito nas computadoras realízase por medio da programación, consistente na acumulación de datos libres de contexto relativos a obxectos, propiedades e accións sobre eses obxectos. O acceso a esa masa de datos é guiado por unhas regras heurísticas que esixen un maior tempo de busca a medida que a cantidade de datos é maior. E aquí é onde parece xurdir unha disparidade entre o funcionamento das computadoras e o comportamento intelixente humano, pois nos seres humanos ocorre á inversa: a medida que dominamos máis un ámbito e nos facemos expertos, podémonos mover con meirande facilidade nel mentres que ás computadoras lles custa máis traballo.

¿A que se debe isto? Aquí é onde entra en xogo un concepto central, *a relevancia*, que depende do contexto. Posto que os programas están formados por datos independentes do contexto, a máquina non pode ter un sentido de que datos son relevantes para unha determinada tarefa, polo que ten que empregar moitos recursos en tatexar heuristicamente na súa base de datos. Así formulado, pode parecer un puro problema tecnolóxico que se podería superar a medida que as técnicas de busca sexan melloradas. Antes ben, o que amosa é que o feito de *estar situados* nun determinado contexto permite aos seres humanos aforrar un enorme esforzo computacional: non necesitamos pensar nin tatexar heuristicamente na nosa mente para recuperar os datos relevantes necesarios para levar a cabo unha tarefa. Eses datos son amosados polo contexto.*

* No capítulo terceiro da primeira sección da obra de Martin Heidegger *Sein und Zeit* ("Ser e tempo"), atopamos un intento de explicación deste proceso. No canto de vivir nun mundo composto de feitos desprovistos de significado (así é como "vive" unha computadora) aos que posteriormente lles asignamos un valor, o ser humano está constitutivamente aberto a (está-en) un mundo de significacións en virtude da súa estrutura fundamental, que é estar-no-mundo. A significación, por tanto, é previa a todo discurso.

Minsky captou a seriedade destas dificultades no que se denominou o problema do sentido común. No noso trato co mundo, os seres humanos manexamos unha gran cantidade de coñecementos implícitos que chamamos o sentido común (por exemplo, coñecementos tan sinxelos como: "Se Ramón está no cuarto do lado, entón o nariz de Ramón está no cuarto do lado."). As computadoras non poden "saber" cousas tan sinxelas coma esta agás que o programador as escriba no seu programa. Imaxinemos agora o problema: ¿que tamaño podería acadar a base de datos do sentido común e de que maneira se pode acceder aos datos relevantes para cada situación? O ser humano xa sabe implicitamente esas cousas porque está *situado en* ou está *aberto a* un contexto no cal hai seres humanos, cada un dos cales posúe o seu nariz —salvo raras excepcións— e os seres humanos atópanse ás veces nun cuarto, co seu nariz incluído. Minsky denominou a cuestión de como as máquinas poden acceder aos datos relevantes para cada tarefa o "problema dos marcos" (*frame problem*), o cal foi amplamente abordado na literatura sobre a IA. Unha maneira de encarar o problema consistiu en tentar dotar os programas

dunhas meta-regras de segunda orde (lembremos que os programas consisten nunhas regras sintácticas —que serían as de primeira orde— para manexar símbolos) que permitirían circunscribir os contextos nos que un conxunto de datos son relevantes. O problema é que esas meta-regras serían máis feitos (accións sobre símbolos) desprovistos de significado, polo que entraríamos nunha regresión infinita de regras para manexar regras.*

En xeral, podemos resumir os problemas aparecidos na IA computacionalista dicindo que o comportamento dos artefactos producidos é totalmente dependente do que o programador poida escribir no seu programa e de que o tipo de escritura da programación estea baseado nesa acumulación de datos libres de contexto, é dicir, universalmente definidos, da que falabamos. A independencia do contexto dos datos dos programas é o que xenerou grandes pro-

* No capítulo 5 da súa obra *What Computers* Still *Can't Do*: *A Critique of Artificial Reason*, Dreyfus expón as dúas posibilidades abertas por esta solución: ou ben caemos no perigo da regresión infinita ou ben teríamos que amosar a existencia duns elementos mínimos libres de toda ambigüidade —libres de contexto— que permitirían funcionar ao programa sen apelar a unhas meta-regras por encima deles. Esta última posibilidade non foi demostrada.

blemas no avance das computadoras, sexan deseñadas para recoñecer caras, traducir idiomas ou para xogar ao xadrez.

Dende mediados da década de 1970, os investigadores trataron de idear artefactos tentando que o papel xogado polo deseñador no seu comportamento fose cada vez menor. Nesta dirección, aproveitaron a aparición a principios da década de 1980 dun novo paradigma nas ciencias cognitivas, o conexionismo, co fin de crear redes neuronais artificiais. Creáronse, por exemplo, algoritmos matemáticos que simulan redes neuronais e que poden aprender e evolucionar —os chamados algoritmos evolucionarios—. Non obstante, este paradigma segue sendo dependente dun concepto que analizaremos a continuación, o de representación. Os anteditos artefactos seguen operando por medio dunha representación simbólica do mundo, polo que o seu desenvolvemento non está baseado nun *trato* co mundo mesmo. Non están, por tanto, *situados nun contexto*.

Iso é o que deu orixe ao último paradigma xurdido nas ciencias cognitivas, a *embodied cognition*, centrado na idea de que a mente, a consciencia ou a intelixencia non poden ser concibidos como algo

que opera independentemente por medio de representacións simbólicas senón que hai que concibilos como algo constitutivamente unido tanto aos nosos corpos físicos coma ao ambiente. Trataremos disto ao final do artigo.

INTELIXENCIA ARTIFICIAL E FILOSOFÍA: ARREDOR DO CONCEPTO DE REPRESENTACIÓN

¿Por que pode ser interesante a IA para a filosofía? Os diferentes proxectos de realización da IA, tanto se fracasan coma se logran levarse a cabo, poden botar luz a cuestións tratadas pola filosofía dende sempre. No caso de que fracasen, amosarían, en principio, que os presupostos que guían aos proxectos, especialmente a súa comprensión do que é o comportamento intelixente humano, son erróneos. No caso de que triunfen, amosarían que é posible simular a intelixencia humana nun artefacto, do cal resultaría unha comprensión da intelixencia mesma, aínda que non hai que desbotar a posibilidade de que unha máquina poida exhibir comportamento intelixente seguindo uns mecanismos diferentes aos utilizados polo ser humano. Por outra parte, posto

que a IA sempre está sometida a consideracións de enxeñaría, podería chegarse á conclusión, no caso dun estancamento severo do proxecto da IA, de que a intelixencia non é reproducible polas técnicas de deseño actuais ou, o que sería máis definitivo e poría de manifesto certos límites da tecnoloxía, que a intelixencia é un atributo exclusivamente humano que non se pode simular en ningún tipo de artefacto —e nese caso tamén obteriamos un entendemento das causas que fan que a intelixencia sexa un atributo exclusivo do ser humano.

Esta interacción entre as teorías sobre a intelixencia e a práctica da IA amósase claramente na íntima relación existente entre as ciencias cognitivas e a IA. A investigación en IA estivo claramente dirixida polos diferentes paradigmas xurdidos nas ciencias cognitivas e, do mesmo xeito, os resultados obtidos na IA inflúen no seguimento ou abandono dun paradigma dentro delas,* xa que constitúen unha maneira útil de probar empiricamente as teorías —aínda

* Como exemplo, Tom Froese expón, no seu artigo "On the role of AI in the ongoing paradigm shift within the cognitive sciences", a influencia que tivo a IA no xurdimento do paradigma da *embodied* e *enactive cognition* no seo das ciencias cognitivas.

que, como apuntabamos antes, esta avaliación empírica é problemática, xa que as posibilidades son diversas: pode ocorrer que unha teoría plausible sobre a intelixencia fracase á hora de aplicarse a unha máquina porque as técnicas dispoñibles non sexan apropiadas, o cal non podería tomarse como unha invalidación definitiva da teoría (este argumento foi esgrimido, por exemplo, polos defensores do paradigma computacional); tamén podería ocorrer que unha determinada aplicación a unha máquina obteña éxitos cuxa extrapolación co fin de comprender a intelixencia humana non sexa lexítima (seguindo a idea exposta máis arriba de que non hai nada que desbote a posibilidade de exhibir un mesmo comportamento intelixente seguindo mecanismos diferentes), o cal poría en dúbida o que anunciamos ao comezo coma a finalidade teorética da IA: a comprensión da intelixencia humana—. Malia a necesidade de ter en conta estes matices, a IA sempre pode botar luz no camiño da comprensión da intelixencia humana.

As relacións entre os investigadores da IA e a filosofía resultan esclarecedoras. Cando Terry Winograd e Hurbert Dreyfus comezaron a introducir a

Heidegger no MIT co fin de amosar un novo camiño aos investigadores, algúns alumnos tiveron reaccións do tipo: "Vós, os filósofos, levades reflexionando nas vosas butacas durante máis de dous mil anos e aínda non entendedes como funciona a mente. Nós no Laboratorio de IA tomamos o mando e estamos triunfando alí onde vós os filósofos fracasastes. Agora estamos programando computadoras para amosar intelixencia humana: para resolver problemas, entender a linguaxe natural, percibir e aprender" (Dreyfus 2, 331).

Atopamos aquí a presuposición de que a IA pode ser entendida á marxe da historia do pensamento filosófico e, o que é máis, en oposición e superación respecto dela. Precisamente é necesario defender o contrario, xa que os presupostos que guían ao paradigma computacionalista dominante en boa parte da IA son dependentes de maneira implícita da conformación subxectivista da metafísica moderna. ¿Que é a intelixencia? Malia que as computadoras son un invento bastante novo, as ideas que guían o paradigma computacionalista teñen moita antigüidade.

Na base da constitución do proxecto metafísico occidental atópase a doutrina substancialista segundo

a cal todo o que é pode ser reducido ou analizado en termos de entes ou substancias (permanentes) e propiedades (variables). Como sabemos, a programación dunha computadora dixital segue ese esquema para poder definir o seu dominio de actuación. A metafísica remítese en última instancia a un principio racional, ou abstracto, que actúa como garantía ontolóxica e epistemolóxica. Na modernidade, atopámonos cunha constitución subxectivista da metafísica xa que é o suxeito quen opera como principio racional último. Partindo da fronteira divisoria na totalidade do ente que establece Descartes entre o mundo físico, ou *res extensa,* e o mundo psíquico, ou *res cogitans*, observamos que o criterio do coñecemento emana do suxeito, do eu que pensa, o cal é entendido dunha maneira formal e universal, atopando a súa expresión máis perfecta nas matemáticas. O coñecemento é a representación por parte do suxeito dun obxecto independente del de acordo a e por medio da legalidade universal expresada nas matemáticas. Preguntabamos antes que é a intelixencia. Pois ben, o computacionalismo é a idea de que a intelixencia humana consiste no manexo dunha representación simbólica do real expresada nunha linguaxe

formal —universal, libre de contexto— cuxos elementos son obxectos e propiedades sobre os que se opera de acordo a unhas regras sintácticas. Expresado na linguaxe da metafísica: é a representación do obxecto seguindo a legalidade universal en que consiste o suxeito. Se as previsións de Minsky, repetidas actualmente con renovada forza, fosen acertadas, se a día de hoxe estivesemos familiarizados con computadoras coma HAL-9000 que dirixisen as empresas ou impartisen clases de filosofía nas universidades, as aspiracións teoréticas da metafísica moderna viríanse cumpridas. Pero seica non foi así.

¿A que se debe a importancia do concepto de representación? A nivel teórico, a idea dunha representación do mundo faise necesaria dende o momento en que impoñemos unha división entre o suxeito e o obxecto e lle concedemos unha primacía á relación teorética entre ambas esferas escindidas. Non ten sentido negar que a capacidade humana de realizar representacións simbólicas do mundo é enormemente frutífera e funciona con éxito en determinados ámbitos. Ora, debemos preguntarmos o seguinte: ¿é posible reducir, dunha maneira extremadamente intelectualista, o compor-

tamento intelixente humano a simples operacións sintácticas sobre representacións simbólicas da realidade? ¿A nosa actividade intelixente consiste nun trato con modelos simbólicos do mundo ou co mundo mesmo?

Estas preguntas están inspiradas nas posturas antirrepresentacionalistas que manteñen, por motivos diferentes, investigadores relacionados coa IA coma Dreyfus ou Rodney Brooks.* Aínda que tal vez non sexa positivo formular a cuestión nuns termos tan definitivos, de xeito que haxa que aceptar ou rexeitar a presencia de representacións, xa que a cuestión empírica sobre a utilización de representacións na cognición humana é mais complexa. Así, malia pretender superar a división entre a mente, o corpo e o ambiente, Andy Clark considera que, dende un punto de vista adaptativo sobre o comportamento humano, o uso de representacións é en cer-

* No seu interesante artigo "Intelligence without representation", Brooks aclara que a súa postura se debe a cuestións de enxeñería exclusivamente e non a unha toma de partido filosófica. Considera que seguir o esquema "percepción (*input*)-representación mental-acción (*output*)" non é adecuado para a creación de robots pois concibe a percepción como algo directamente ligado á acción sen que medie un centro de control executivo.

tos casos útil e necesario para a economía cognitiva humana. Pero o que debemos pór en dúbida é se é posible *basear* a cognición na representación. As aspiracións da IA computacionalista dependen dunha resposta afirmativa a esta cuestión, é dicir, da posibilidade de concibir a cognición como o manexo de representacións simbólicas do mundo que descansan, en última instancia, nun conxunto de feitos e regras expresados nunha linguaxe formal e universal independente do contexto, libre de toda ambigüidade e, en consecuencia, de toda interpretación. Ora, ¿por que consideramos que estas aspiracións non se poden cumprir?

Xogar ao tenis con destreza é unha actividade intelixente. Implica gran desenvolvemento dunhas capacidades sensomotoras que conectan a percepción do xogo coas reaccións musculares do corpo, isto é, unhas capacidades cognitivas que permiten entender e avaliar o xogo, darlle ordes ao corpo, decidir como, cando e cara a onde debe moverse, "pensar" nas estratexias apropiadas diante do estilo de cada contrincante. Cando unha tenista está recibindo unha pelota a 120 km/h, non realiza unha representación simbólica do mundo en forma de

ecuacións diferenciais cuxa resolución lle sinalaría o momento no que debe estar preparada para recibila e golpeala; unha tenista non está resolvendo ecuacións mentres xoga. O que fai ou, mellor dito, *o que fixo*, é interactuar repetidamente *de maneira física* cun ambiente no que o seu corpo recibe pelotas a gran velocidade, aprendendo a responder *corporalmente* de maneira axeitada a estas situacións tantas veces repetidas. *Fóra do xogo*, as representacións físico-matemáticas que pode empregar unha adestradora como parte da aprendizaxe poden resultar útiles para comprender a traxectoria das bólas ou a velocidade de impacto; pero esa comprensión non pode substituír, e moito menos basear, a aprendizaxe corporal que se produce na pista, xogo tras xogo. "Pensar co corpo" non é unha metáfora, xa que realmente é iso o que acontece: a cognición é unha actividade máis ou menos complexa de conexión, que evidentemente pode recorrer ao simbolismo, entre a percepción e a acción. Pensar outra cousa, crer que a cognición é unha dimensión "abstracta" que non ten nada que ver coa actividade dun corpo que percibe, sente e actúa é esquecer unha cousa moi básica: somos seres vivos que temos que actuar no mundo.

Poñendo outros exemplos, a posibilidade da comprensión do sentido dunha frase, da función dun martelo ou a diagnose dunha enfermidade a partir duns síntomas descansa nunha precomprensión implícita deses fenómenos que non é verbalizable na súa totalidade. Esa precomprensión constitúe un *background* de asuncións implícitas que non podemos reducir a un conxunto de feitos e regras independentes do contexto. Expresándoo dun xeito máis sinxelo, a destreza humana, o seu saber-facer, non pode ser recollida nun manual. Non pode existir o manual da boa filósofa, do bo médico ou do bo músico. Por iso constitúe unha pretensión fabulosa o intento de trasladar a intelixencia humana a unha computadora dixital. Podemos comprender o mundo porque estamos situados nel —é dicir, porque estamos existencialmente abertos ás súas significacións— e non porque poidamos facer representacións del. Por tanto, as características da actividade intelixente, a cal presupón poder estar situados nun contexto equipados cunha sabedoría implícita que nos permite *interpretar* os fenómenos, parecen levar á superación da creación de IA baseada en presupostos computacionalistas, ou exclusivamente representa-

cionalistas. Eses intentos de superación son os que exporemos a continuación.

AS NOVAS CORRENTES DAS CIENCIAS COGNITIVAS: *EMBODIED* E *ENACTIVE COGNITION*

Segundo a perspectiva aquí seguida, a historia da metafísica constitúe un lugar de explicación adecuado dende o cal se poden comprender os fundamentos e as limitacións dos diversos paradigmas que operan no proxecto de construción da IA. Así, puidemos comprobar que os presupostos que alimentan o primeiro paradigma dominante nas ciencias cognitivas, o cumputacionalismo —e tamén, nun certo senso, o segundo paradigma: o conexionismo— encaixan coas aspiracións subxectivistas da metafísica moderna. Dende o século XIX, esta metafísica, asentada nunha división entre o suxeito e o obxecto e nunha comprensión intelectualista do suxeito pola cal o coñecemento que este poida ter do obxecto é entendido en termos de representacións simbólicas, formais e universais cuxa forma máis perfecta é a representación matemática, sofre, coma ben é sabido, unha crise.

Para poñer a proba esta hipótese, segundo a cal é posible comprender a problemática suscitada pola IA atendendo aos presupostos metafísicos presentes nela, podemos comprobar en que medida as novas propostas metafísicas xurdidas a raíz da antedita crise se atopan presentes ou resoan nos intentos de creación de IA que someten a crítica as aspiracións do computacionalismo, xa que estes intentos, como veremos, non son asimilables a unha metafísica intelectualista da subxectividade. E o xurdimento nas últimas tres décadas dun novo paradigma no seo das ciencias cognitivas que tenta facer fronte ao estancamento da IA baseada en presupostos computacionalistas, denominado *embodied cognition,* ou poscognitivismo, parece confirmar a hipótese.

De entrada, é necesario aclarar que non se pode facer unha exposición lineal na que unha determinada perspectiva ou paradigma metafísico supostamente erróneo é substituído por outro paradigma que finalmente é considerado o correcto, xa que a corrección non é unha categoría útil para orientarse en cuestións metafísicas. A cuestión é apreciar rigorosamente —que non correctamente, a diferenza é importante— o poder explicativo que posúen as

creacións especulativas. Se aquí facemos unha crítica da IA computacionalista e dos presupostos metafísicos que subxacen nela, non é tanto para defender que as súas propostas son incorrectas como para amosar que se trata dun poder que tentou exceder os seus límites. Isto explica, por unha parte, os éxitos que logrou o computacionalismo en determinados ámbitos e, por outra, o relativo fracaso das súas desmesuradas aspiracións. É esa extralimitación a que tamén explica e xustifica a aparición do poscognitivismo. A idea aquí defendida é que a intelixencia non pode ser reducida a operacións sintácticas efectuadas sobre representacións simbólicas do real; malia que certos procesos cognitivos se poden comprender seguindo ese esquema, o que se deixa de lado nel impide acadar unha comprensión integral da intelixencia. E o paradigma dunha cognición corporal ou incorporada, dunha *embodied cognition*, é interesante porque revela que é o que se deixou de lado.

SUXEITOS CON CORPO: O REXEITAMENTO DO INTELECTUALISMO

Unha constante que atopamos ao longo da historia da metafísica é o rexeitamento do corpo e, en xeral, de toda realidade material. Tanto na súa primitiva constitución ontoteolóxica coma na moderna constitución subxectivista, o principio ontolóxico último é concibido como unha realidade inmaterial, abstracta ou intelectual. O suxeito moderno non consiste nos diversos suxeitos empíricos senón que é un suxeito formal, universal e, podemos engadir, ahistórico: a razón. É, por tanto, un suxeito *sen corpo, sen afeccións* e *sen emocións*; o pensamento é concibido coma un proceso puramente intelectual que está desligado da nosa constitución como seres biolóxicos. Non hai continuidade entre bioloxía e psiquismo, o que explica o antropocentrismo presente en boa parte das propostas metafísicas modernas: os seres vivos son simples corpos determinados mecanicamente polos seus instintos, do mesmo xeito que se transmiten as forzas nas engrenaxes dun reloxo; o ser humano, pola súa parte, é un ser dotado de razón e autodeterminado por ela.

Nietzsche introduce unha creba na historia da metafísica ao poñer no centro da súa filosofía o con-

cepto de corpo. O pensamento xa non pode ser entendido á marxe da nosa constitución como seres biolóxicos cunhas necesidades e uns obxectivos. Iso explica a constante aparición na obra de Nietzsche da cuestión dos instintos; a repetida afirmación, a modo de recordatorio, de que o ser humano é *un animal* e, en xeral, a articulación que tenta establecer entre a actividade intelectual e os procesos fisiolóxicos. Malia que a súa filosofía non é recoñecida polo paradigma da *embodied cognition* como unha influencia directa, pódese observar que estas preocupacións están na base da súa constitución, tal como o seu nome indica.* Fronte á concepción da percepción como representación e da cognición como computación propia do computacionalismo, a denominación deste paradigma sinala a importancia que posúe, nos procesos cognitivos, o feito de estar situados *corporalmente* nun ambiente co que interactuamos continuamente, polo que a férrea división entre

* Non resulta doado atopar unha tradución do termo *embodied* que recolla a idea que pretende transmitir. Posibles solucións son "encarnado", "incorporado" e "corporeizado" pero ningunha delas nos parece satisfactoria; por iso preferimos deixalo en inglés. *Enactive*, pola súa parte, é perfectamente traducible coma "enactivo".

o corpo, a mente e o ambiente deixa de ser válida. A cognición concíbese como un fenómeno biolóxico no que o organismo é visto como un sistema dinámico que se autoorganiza en virtude da súa indisoluble unión co ambiente.

Esta apreciación dunha continuidade entre vida, mente e cognición está asentada na teoría biolóxica da autopoiese desenvolvida por Humberto Maturana e Francisco Varela, a cal ten a súa orixe, entre outros elementos, nas investigacións do propio Maturana sobre a percepción das ras. Nos seus estudos, Maturana atópase con que non podemos seguir falando de sistemas que perciben unha realidade externa obxectiva e independente do observador. Non hai unha conexión directa entre as lonxitudes de onda que afectan a retina da ra e as percepcións xeradas no seu sistema nervioso. Máis ben son o resultado das relacións entre os sistemas neuronais concibidos como un todo. As perturbacións provocadas polo exterior alteran as neuronas pero é a estructura destas últimas a que xera as percepcións: "As perturbacións non determinan o que ocorre no sistema nervioso senón que simplemente provocan cambios de estado. É a estrutura do sistema pertur-

bado a que determina ou, mellor dito, *especifica* que configuración estrutural do medio pode perturbalo" (Flores e Winograd, 43).

Partindo destas ideas, Maturana e Varela desenvolveron a teoría biolóxica da autopoiese. Os organismos son comprendidos como entidades que acadan a autonomía por medio da autoconstitución, baixo unhas condicións precarias, das súas estruturas, caracterizadas por unha clausura organizacional. As estruturas dos seres vivos xeran e destrúen os seus compoñentes, así como os propios procesos de produción dos mesmos, para manter o seu equilibrio homeoestático. Hai unha unión estrutural co ambiente na cal o sistema persigue a fin de non desintegrarse (de aí que se fale dunha clausura organizacional) por medio de reaxustes estruturais que non poden ser concibidos como simples representacións dun mundo externo. A continua autoconstitución do sistema, que é o que asegura a súa supervivencia, é provocada pola súa unión estrutural co ambiente. Estas ideas constitúen un rexeitamento da visión condutista na que se reduce a conduta á pura reacción ante uns estímulos externos. Dáselle unha primacía ás estruturas internas do sistema que

son as que, como diciamos, especifican cales estímulos poden afectarlle e se autoestruturan en base aos estímulos recibidos, polo que a división interior-exterior, organismo-ambiente, se esvaece, malia que iso non impide darlle unha grande importancia a antedita autoestruturación (que sería a parte interna do proceso pero neste caso só cobra sentido pola súa unión co exterior). A cognición xa non pode ser considerada como o manexo por parte dunha mente independente do corpo e do ambiente por medio de representacións simbólicas dunha realidade externa, independente e obxectiva, senón que consiste nunha autoconstitución de estruturas que están unidas ao seu ambiente.

Iso lévanos a apreciación dunha das grandes influencias presentes no paradigma que nos ocupa, influencia que é neste caso recoñecida explicitamente. A filosofía fenomenolóxica e, en especial, o pensamento de Heidegger, constitúen a súa principal referencia metafísica e por iso Dreyfus chega a falar dunha IA heideggeriana. Malia que se podería sinalar o antropocentrismo propio da metafísica moderna no privilexio ontolóxico que Heidegger lle concede ao *Dasein*, a análise deste levada a cabo en *Sein und*

Zeit é aproveitada como un punto de partida para unha explicación non intelectualista da cognición. A estrutura fundamental do *Dasein* como ser-no-mundo aspira ao quebrantamento da división suxeito-obxecto propia da modernidade. Segundo unha interpretación pragmatista do pensamento de Heidegger, esta estrutura amosaríanos que a nosa relación primaria co mundo non é unha relación teorética na que operariamos por medio de representacións simbólicas do mundo senón que é unha relación práctica na cal a actividade do suxeito e a significación do obxecto están esencialmente unidos: estamos existencialmente abertos a un mundo de significacións. Unicamente cando se produce unha avaría, dinos Heidegger, un fallo nesa relación práctica, é cando o obxecto se tematiza expresamente, dirixindo así a nosa atención cara a el e marcando unha distancia teorética cara ao mesmo. O saber-tratar, o saber-facer ou o saber-como anteceden ontoloxicamente ao saber-que.

En consecuencia, o proceso da actividade intelixente non consiste nunha descomposición da realidade en unidades ás que lles atribuímos unha representación simbólica para poder realizar cálculos

por medio dos cales obteriamos a resposta desexada, é dicir, a actividade buscada. Ou, mellor dito, a actividade intelixente non pode ser reducida completamente a ese proceso. Os procesos intelectuais abstractos, desligados da interacción física co ambiente, xogan un certo papel nas tarefas cognitivas e poden ser máis ou menos prominentes segundo a natureza da tarefa. Pero esas tarefas cognitivas non se poden desligar da nosa constitución como seres biolóxicos cunhas necesidades e obxectivos que dotan de significación o ambiente nin poden ser comprendidas á marxe da interacción con este último; así, autores como Lakoff tentan amosar que incluso os pensamentos máis abstractos están relacionados con metáforas espaciais baseadas nesa interacción. ¿Por que situamos a Dios nos ceos ou falamos dos "baixos" instintos? ¿Por que hai persoas "sinistras" ou falamos de "elevadas destrezas"? ¿Non procederá dunha asignación *biolóxica* de valor á altitude, dunha identificación da intelixencia co cerebro que se sitúa no alto do corpo e dun privilexio dos seres vivos destros en detrimento do baixo e dos zurdos? Lembremos a vella alma tripartita dos gregos: no baixo ventre atópanse os instintos, na parte

media do corazón a vontade e na cabeza o máis "elevado", a alma racional.

Podemos acudir a outros exemplos para ilustrar o carácter corporal da cognición. Na película *Matrix* hai unha escena na que se plasma a esencia das vellas aspiracións do computacionalismo. Nela, a protagonista Trinity necesita aprender a pilotar un helicóptero e solicita que carguen, debemos supoñer que no seu cerebro, o programa informático que lle permita facelo. Así, presuponse que o saber-facer, a actividade intelixente en que consiste pilotar un helicóptero, é reducible a un conxunto de regras abstractas aplicadas a un conxunto de símbolos. E isto é exactamente o que consideramos que non é posible. Para aprender a pilotar un helicóptero, *hai que pilotalo*; pilotar inclúe a interacción física co aparato, o manexo dos seus elementos e obriga tamén a ter en conta a interacción do aparato co seu entorno en diferentes condicións (vento, visibilidade, etc.), o que implica o desenvolvemento dunha destreza práctica na que a cognición máis ou menos abstracta e a interacción corporal van da man. Certos coñecementos abstractos son, por suposto, necesarios pero en ningún caso poden ser considerados suficientes.

Trasladándonos a un terreo máis actual —as interesantes especulacións de *Matrix* datan xa do 1999—, poderíase pensar que o desenvolvemento dos coches autónomos —e, en xeral, o *boom* do automatismo ao que asistimos hoxe en día— invalida o noso argumento. E aquí non estamos falando de especulacións senón dunha realidade efectiva: tras anos de ensaios e probas nos que destaca o coche creado por Google, o 7 de novembro de 2017 a compañía Waymo anunciou que os seus coches totalmente autónomos xa están dispoñibles para aqueles voluntarios que queiran facer probas reais de tráfico con eles co fin de establecer unha flota de taxis autónomos no estado de Arizona, pioneiro na autorización dese tipo de vehículos. Pero ese automatismo exemplificado nos coches non debe ser comprendido como un triunfo da redución da actividade intelixente e da cognición a unha computación abstracta. Máis ben, o que amosa é a necesidade da inclusión de diversos sistemas sensomotores para poder realizar unha actividade intelixente como é conducir: o sistema LIDAR, unha sorte de radar de luz que, mediante láser, crea un mapa do contorno de xeito similar a como o fan os morcegos mediante ultrasóns; radares habituais de

ondas de radio que permiten detectar obstáculos; sistemas de posición GPS, cámaras, sensores de medición inercial, etc. En resumo, todo un conxunto de sistemas de percepción e interacción co ambiente que son asimilables á actividade sensomotora dun ser vivo. *Agás nunha cousa*: a necesidade de computación dos coches para manexar os datos percibidos e acumulados é enorme, ata o punto de que a computadora que leva o coche non é suficiente e necesita conectarse pola Internet para obter o soporte computacional óptimo. O condutor humano non necesita ese nivel de computación xa que a súa condución está asentada nun *background* de innumerables interaccións físicas co coche e co ambiente, as cales están *incorporadas* á súa actividade sensomotora.

En calquera caso, o coche autónomo resulta interesante porque é, dalgún xeito, a realización efectiva dun híbrido de ser vivo e da IA computacionalista. Recolle así, en parte, as premisas da elaboración de IA a partir do paradigma *embodied*, o que pon o acento na necesidade de enfrontar o problema da situacionalidade (*situatedness*), é dicir, considerar como requisito da intelixencia estar situados no mundo; é ese estar xa sempre situados o que con-

fire unha relevancia, unha significatividade aos elementos do ambiente. Por iso a IA non se pode limitar á aplicación de programas informáticos nunha máquina; en troques de que o deseñador determine por completo e de xeito externo o comportamento da máquina, esta ten que posuír un corpo que perciba e interactúe co ambiente co fin de desenvolver, por medio da aprendizaxe, as destrezas prácticas que lle permitan alcanzar os obxectivos buscados. Así, creouse IA *embodied* incluíndo circuítos sensomotores pechados que son sensibles ao contexto combinados con algoritmos evolucionarios. Non obstante, a creación de IA baseada nas ideas *embodied* topou tamén con dificultades, tal como expón Dreyfus. Eses son os problemas que tentan ser encarados pola radicalización da corrente *embodied* levada a cabo polo enactivismo.

COGNICIÓN E EMERXENCIA: UNHA ALTERNATIVA AO SUBXECTIVISMO REPRESENTACIONALISTA

A imposición dunha férrea división entre suxeito e obxecto implica unhas graves consecuencias que foron minimamente sinaladas ao longo do arti-

go. A nivel ontolóxico, establécese unha distancia insuperable entre o pensamento e a materia, así como entre a constitución biolóxica e a realidade psíquica. A imposibilidade de apreciar unha continuidade entre as diversas esferas do real deixa como unha única vía posible a redución dunhas a outras; ironicamente, esta redución realízase en ambas as direccións: o obxecto é reducido ás categorías intelectuais do suxeito e, ao mesmo tempo, o pensamento do suxeito, a súa realidade psíquica, é reducido ao correlato material no que se supón que ten asento: o cerebro. Así, nun único movemento, *a materia é convertida en algo abstracto e o pensamento é corporeizado en algo material*.

O concepto de emerxencia ofrece unha vía de escape na que a redución é substituída pola articulación. O necesario recoñecemento da diferenza, da pluralidade das diversas esferas da realidade (materia e vida, vida e pensamento, organismo e ambiente, etc.) non implica, neste caso, o establecemento dunha división entre elas nin a redución dunhas a outras: a apreciación dunha continuidade non está enfrontada co recoñecemento da pluralidade. As propiedades emerxentes dun sistema son propieda-

des que non son reducibles ás propiedades dos compoñentes do sistema, é dicir, a interacción entre os compoñentes do sistema nun nivel micro dan lugar á aparición de propiedades novas nun nivel macro. De aí que o concepto de emerxencia permita dar conta da articulación entre diversos elementos sen reducir uns elementos a outros, xa que aparecen novos niveis de organización que engloban os anteriores.

A cognición pode ser entendida como un deses niveis de organización que emerxen a partir da interacción entre diversas esferas. Iso apréciase claramente noutro dos elementos recollidos polo paradigma *embodied*: a teoría da mente estendida desenvolvida por Andy Clark e David Chalmers.* Nela, a mente non se entende como algo limitado pola "carne e o cranio" senón como o resultado da interacción presente no *continuum* ontolóxico mente-corpo-ambiente. Colle forza a idea dun sistema que engloba ese *continuum* no que as propiedades emer-

* Pódese rastrexar a orixe da teoría no artigo "The Extended Mind", de Clark e Chalmers; para unha exposición máis pormenorizada destas ideas, véxase o libro de Clark titulado *Being There: Putting brain, body, and world together again*.

xentes posúen unha grande importancia e a idea de que unha das esferas do sistema —a mente— realiza representacións doutra —o mundo— perde pulo. As destrezas humanas son concibidas como a aparición, sen que haxa un plan racional previo por parte dun centro organizador, de propiedades emerxentes no sistema mente-corpo-entorno. E a emerxencia non permite ser entendida en base á descomposición do sistema nos seus elementos constituíntes, o que explica que a división entre estes elementos, propia da metafísica subxectivista, non sexa útil. Antes ben, a emerxencia consiste na aparición de variables incontrolables que xorden ou emerxen a partir da interacción entre os compoñentes do sistema.

Neste punto debemos deternos e observar o perigo que supón, tal como nos alerta Isabelle Stengers,* converter a emerxencia nun concepto todoterreo, pois poderiamos caer na mesma extralimitación de poder que detectamos previamente no computacionalismo. A pesar da revalorización do concepto de emerxencia presente nos estudos sobre os sistemas complexos, é un concepto que permanece en certa

* Véxase o seu *Cosmopolitiques VI: La vie et l'artifice: visages de l'émergence*, en Isabelle Stengers, *Cosmopolitiques*, 2 vols.

medida en estado especulativo e os intentos por transformalo nun programa de investigación científico están tendo lugar de xeito bastante recente.* Aquí apreciamos que a IA pode xogar un papel especial xa que, en consonancia co que vimos defendendo, a IA é unha *práctica* (é dicir, unha efectiva construción empírica que pode funcionar ou non) que está, dada a natureza das súas aspiracións, intrinsecamente conectada coa especulación metafísica. O que non quere dicir, en ningún caso, que quen elaborou a IA teña que compartir estas preocupacións metafísicas (tal como declara Brooks), ou que o éxito dun determinado programa de IA poida ser tomado como unha validación empírica de certa perspectiva metafísica. Ningunha especulación metafísica pode ser validada empiricamente; ao contrario, son os programas de investigación empíricos os que asumen certos presupostos metafísicos que non se poden demostrados.

A IA ocupa un lugar especial porque ofrece un terreo de experimentación no que se pon a proba o

* Para un desenvolvemento desta tese, véxase o artigo de Claus Emmeche *et al.*, "Explaining emergence: towards an ontology of levels".

alcance do poder dos conceptos, realizando un movemento no que se abandona o terreo especulativo e se inicia a tarefa de construír artefactos que pretenden realizar actividades propias dun comportamento intelixente. Nese sentido, podemos observar a importancia que adquire o concepto de emerxencia nas ideas enactivas. O enactivismo radicaliza a idea, propia dos intentos dinámicos e evolucionarios da IA *embodied*, de que o deseñador debe intervir o menos posible no artefacto. Considera que non é suficiente con dotar o axente cognitivo dun corpo que interactúe co ambiente e dunhas ferramentas que lle permitan aprender e evolucionar para que poidamos falar de axentes cunha perspectiva significativa. A relación de asignar relevancias non debe ser realizada no axente artificial senón que debe xurdir espontaneamente neste e, para iso, debe estar concernido pola súa propia existencia, é dicir, debe ter un si-mesmo que actúe cunha finalidade propia.

Isto lévanos a unha idea central da corrente enactiva: a autonomía constitutiva. Seguindo a concepción kantiana de propósito natural, considérase que os seres vivos posúen unha teleoloxía inmanente

en virtude da cal se autoproducen ou se autoconstitúen. Os seres vivos, desenvolvidos nun ambiente de condicións precarias que ameazan a súa existencia, dótanse a si mesmos dunha finalidade. A autonomía constitutiva, isto é, a capacidade por parte do axente de autoconstituír as súas estruturas dotándose así dunha identidade, diferenciándose do ambiente e non desintegrándose, é o que permite o mantemento dunha finalidade *propia*.

En lugar de dotar o axente cun sistema de valores para dar significado ao mundo, o enfoque enactivo tenta crear as condicións de emerxencia para que o axente se autoconstitúa, dotándose a si mesmo dese sistema de valores. As diferenzas coa IA computacionalista son notables. Non se crea un mundo, en forma de representación simbólica, para aplicalo no axente por medio dun programa informático. Ao contrario, considérase que é o axente quen debe enactuar o seu mundo, isto é, crear unha rede de significacións en base ao que é relevante para a súa propia existencia. Para posuír un mundo, hai que crealo; e unicamente creamos se estamos concernidos por algo. Non deberiamos pensar, non obstante, que esta creación de mundo é a actividade dun

suxeito escindido do seu ambiente; máis ben, é algo que emerxe froito da interacción entre un axente cuxo interese propio é o autosostemento e un ambiente que á vez posibilita e pon límites á súa existencia.

Bibliografía

Brooks, Rodney A.
"Intelligence without representation", *Artificial Intelligence*, nº 47, 1991, páxs. 139-159.

Clark, Andy
Being There: Putting brain, body, and world together again, Cambridge: The MIT Press, 1997.

Clark, Andy e Chalmers, David J.
"The Extended Mind", *Analysis*, nº 58, 1998, páxs. 10-23.

Dreyfus, Hubert L.
1. *What Computers* Still *Can't Do: A Critique of Artificial Reason*, Cambridge: The MIT Press, 1992.
2. "Why Heideggerian AI failed and how fixing it would require making it more Heidegger-

ian", in: Philip Husbands *et al.* (eds.), *The Mechanical Mind in History*, Cambridge: The MIT Press, 2008, páxs. 331-371.

Emmeche, Claus *et al.*
"Explaining emergence: towards an ontology of levels", *Journal for General Philosophy of Science*, nº 28, 1997, páxs. 83-119.

Flores, Fernando eWinograd, Terry
Understanding computers and cognition. A new foundation for design, Norwood: Ablex Corporation, 1990.

Floridi, Luciano
"Should we be afraid of AI?", *Aeon*, 9/5/2016. URL: <https://aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible>

Froese, Tom
"On the role of AI in the ongoing paradigm shift within the cognitive sciences", en: Max Lungarella *et al.* (eds.), *50 Years of AI*, Berlín: Springer-Verlag, 2007, páxs. 63-75.

Heidegger, Martin
Ser y tiempo, Madrid: Ed. Trotta, 2009.

Maturana, Humberto e Varela, Francisco
De máquinas y seres vivos. Autopoiesis: *la organización de lo vivo*, Santiago de Chile: Editorial Universitaria, 1998.

Stengers, Isabelle
Cosmopolitiques, *2 vols.*, París: La Découverte, 2003.

Turing, Alan M.
"Computer Machinery and Intelligence", *Mind*, nº 49, 1950, páxs. 433-460.

Índice

EDICIÓN NON VENAL
ISSN 2340-8537