EL PROBLEMA DEL MARC I ELS AE's

TREBALL DE FINAL DE GRAU

Alumne: Jordi Aparicio Llorens

Tutor: David Pineda Oliva

Curs acadèmic 2019/2020

Universitat de Girona

Facultat de Lletres

Doble Grau en Economia i Filosofia

Índex

Resum

El problema del marc va ser detectat primerament en l'àmbit de la intel·ligència artificial (IA) i va passar a tenir repercussions en l'àmbit de la filosofia de la ment. Aquest problema va generar un intens debat entre filòsofs i experts en IA sobre què s'havia d'entendre per 'problema del marc', així com diverses temptatives de solució al suposat problema. En aquest treball, tractaré la visió de Daniel Dennett sobre el problema del marc i proposaré una possible solució: els algoritmes evolutius. Discutiré els punts a favor i en contra de la meva proposta i diré en quins aspectes pot solucionar el problema del marc i acostar la IA a la intel·ligència humana.

Paraules clau: problema del marc, IA, selecció natural, AE's, Dennett.

1. Introducció

L'objectiu d'aquest treball és entendre la visió de Daniel Dennett sobre el problema del marc i discutir si els algoritmes evolutius representen una solució al problema. Les implicacions filosòfiques d'aquesta mena d'algoritmes no han sigut prou estudiades en filosofia (Muntean 2014), però crec que el seu estudi pot contribuir tant a solucionar una concepció particular del problema del marc com a entendre la ment humana.

No hi ha un consens a l'hora de definir què s'ha d'entendre per 'problema del marc' dins de la comunitat filosòfica ni tampoc dins de la comunitat d'experts en intel·ligència artificial (IA). En aquest treball, no tractaré definicions alternatives del problema del marc per part d'altres autors que no siguin Dennett ni em centraré en la discussió sobre què engloba aquest problema. Tampoc em centraré en el debat de com reduir el problema del marc a d'altres problemes (problema de la rellevància inductiva, problema del coneixement implícit, problema de la inferència demostrativa, etc.) (Pylyshyn 1987).

El debat sobre la definició d'aquest problema obre la porta a un estudi prou exhaustiu. Per aquest motiu, em limitaré a comentar la visió de Dennett i no en comentaré d'altres com bé podrien ser la dels filòsofs Fodor i Haugeland o la dels experts en IA: Hayes i McDermott, entre d'altres (Pylyshyn 1987). Tractar altres visions sobre el problema del marc faria més interessant, però també més dens, el meu treball.

La tesi principal del meu treball és que els algoritmes evolutius ofereixen una resposta al problema del marc segons l'entén Dennett. Un camp en el que s'han estudiat les implicacions filosòfiques d'aquests algoritmes és el de la filosofia de la ciència. Alguns autors han estudiat les semblances dels algoritmes evolutius (descrits com a *machine learning programs* en la literatura) amb la filosofia de la ciència (Casacuberta & Vallverdú 2019; Korb 2004). Aquests autors assenyalen que la filosofia de la ciència i els algoritmes evolutius comparteixen certs interessos: els dos pretenen descobrir lleis bàsiques, els dos es regeixen per principis de simplicitat, etc. Però les diferències també són considerables. Els algoritmes evolutius pretenen solucionar problemes concrets i són més pràctics que teòrics. En canvi, la filosofia de la ciència és més teòrica que pràctica: s'interessa per l'anàlisi conceptual dels termes científics, pretén explicar el mètode científic, etc. Jo em centraré en estudiar les implicacions d'aquests algoritmes pel que fa el problema del marc, seguint amb la línia del que apunten alguns autors:

> NARS (Non-Axiomatic Reasoning System) or NAL (Non-Axiomatic Logic) is based on these new statistical approaches, as well as the result of the implementation of non-monotonic logics and reasoning to several domains of academic research, making possible the study or more complex problems (Korb, 1992) and to find solutions to classic ones, like the frame problem. (Casacuberta & Vallverdú 2019)

A l'apartat 2, faré un breu repàs de l'origen del problema del marc i definiré aquest problema tal com l'entén Dennett mitjançant un exemple seu. A l'apartat 3, presentaré els algoritmes evolutius a través d'una descripció tècnica general sobre el seu funcionament, la seva estructura i d'un exemple. A l'apartat 4, argumentaré que els algoritmes evolutius representen una solució al problema del marc descrit per Dennett. Finalment, a l'apartat 5, resumiré les conclusions principals del treball i apuntaré noves línies de recerca possibles inspirades en el meu treball.

2. Problema del marc

Un dels problemes en IA és el problema del marc (*the frame problem*). Abans de tractar les implicacions filosòfiques que Dennett detecta en el problema del marc, faré un breu repàs del seu origen. Crec que l'entrada de l'*Standford Encyclopedia of Philosophy* resumeix perfectament els dos enfocaments teòrics (el lògic i el filosòfic) que tracten d'esbrinar què és el problema del marc:

> To most AI researchers, the frame problem is the challenge of representing the effects of action in logic without having to represent explicitly a large number of intuitively obvious non-effects. But to many philosophers, the AI researchers' frame problem is suggestive of wider epistemological issues. Is it possible, in principle, to limit the scope of the reasóning required to derive the consequences of an action? And, more generally, how do we account for our apparent ability to make decisions on the basis only of what is relevant to an ongoing situation without having explicitly to consider all that is not relevant? (Standford Encyclopedia of Philosophy)

El text anterior ja palesa que el problema del marc és entès diferentment segons si s'estudia des de l'òptica de la IA o si es fa des de la filosofia. Pels experts en IA, el problema del marc representa un problema lògic: és el problema de dir quins fets no varien en esdevenir-se una acció sense haver d'explicitar un gran nombre (virtualment infinit) de fets obvis que no varien (Kamermans & Schmits 2004). En canvi, pels filòsofs, el problema del marc implica problemes epistemològics, alguns d'ells els comentaré quan parli de la visió de Dennett.

2.1. Origen del problema del marc

El problema del marc va ser detectat originalment per McCarthy i Hayes l'any 1969 com un problema lògic concernent a la IA. A grans trets, el problema del marc, segons McCarthy i Hayes, té a veure amb la manca de coneixement sobre què roman igual a mesura que el món canvia:

> [...] the Frame Problem is really not much more than the problem of finding a logical representation that adequately describes what doesn't change when actions take place in a dynamical world. (Kamermans & Schmits 2004)

En un món en el que mires el rellotge i no està especificat que mirar el rellotge no impliqui canviar el color dels teus mitjons, no exclou la possibilitat que els teus mitjons canviïn de color cada cop que miris el rellotge. Sembla que la solució trivial d'un problema com aquest és explicitar que observar el rellotge no altera el color dels teus mitjons. Tanmateix, això ens portaria a haver de fer explícites totes les accions que no alteren cap altra part del món, de manera que podem imaginar infinits enunciats de no-implicació possibles que es podrien formalitzar per descriure el nostre món ordinari.

A la realitat, és obvi que quan mires el rellotge, el color dels teus mitjons no canvia. Però això no és del tot clar en un sistema basat en IA. Per tal de garantir que certs fets estranys no s'esdevinguin en un sistema d'IA, cal introduir certs axiomes. Els axiomes que introduïm en IA per tal de garantir la persistència de certs objectes o propietats del món són els anomenats *axiomes del marc*. Aquests axiomes no poden ser violats per cap acció que es dugui a terme en el món en qüestió. El nombre d'axiomes del marc necessaris per descriure adequadament les situacions és considerablement gran en comparació amb el domini d'objectes, propietats i accions; en general, la majoria d'objectes i propietats del món no són afectats per les accions. Generalment, cada cop que afegim una acció nova, haurem d'afegir tants axiomes del marc com objectes i propietats hi hagi, i cada cop que afegim un objecte o propietat nou, haurem d'afegir tants axiomes del marc com accions. El nombre total d'axiomes del marc requerits per un domini de n objectes o propietats i m accions serà de $m \times n$.

Aquest problema de la IA, detectat inicialment com un problema lògic, es trasllada a l'àmbit de la filosofia com un problema de coneixement. El problema del marc s'ha descrit de diferents maneres en l'àmbit de la filosofia. Fodor, per exemple, l'ha reformulat de la següent manera:

> [...] the Frame Problem is just [the problem when to stop thinking] from an engineer's perspective. (Fodor 1987)

En general, no hi ha un consens a l'hora de definir el problema del marc (Pylyshyn 1987); totes les descripcions sobre el problema del marc són diferents en algun aspecte. Dennett i Fodor creuen que el problema del marc, a més de ser un problema concernent a modelar el món en termes lògics, implica un problema epistemològic fonamental.

2.2. Dennett i el problema del marc

La primera referència al problema del marc des d'un punt de vista filosòfic es pot trobar a Dennett (1984), en el seu article *Cognitive Wheels: The Frame Problem of AI*. En aquest article, descriu el problema del marc de la següent manera:

> […] a new, deep epistemological problem –accessible in principle, but unnoticed by generations of philosophers– brought to the light by the novel methods of AI, but still far from being solved. (Dennett 1984)

Per explicar el problema del marc, Dennett descriu dues situacions de robots que fracassen en intentar complir les seves respectives tasques per culpa, justament, del problema del marc. Comentaré un exemple de Dennett perquè em servirà per explicar el funcionament dels algoritmes descrits a l'apartat 3.2., els quals presentaré com a proposta de solució del problema del marc.

Dennett entén el problema del marc com la incapacitat de descartar, en contextos ordinaris, la informació irrellevant amb la mateixa agilitat que un humà. Per *informació rellevant* consideraré tota aquella informació que pot contribuir significativament a realitzar una tasca determinada. Amb *informació irrellevant* em referiré a tota aquella informació que no contribueix en res a solucionar una tasca determinada i que, per aquest motiu, ha de ser descartada. Segons Dennett, el problema del marc equivaldria al problema de la rellevància:

> El problema es, así pues, el siguiente: cuando ejecutamos una acción, la mayor parte de información nueva que nos llega es irrelevante para el fin que perseguimos con la acción y por tanto, podemos ignorarla por completo [...], pero cuando ocurre algo relevante, debemos atender a esa circunstancia y ajustar nuestra conducta para que, en la nueva circunstancia, logremos el fin perseguido, y debemos además actuar con suma rapidez y diligencia. (Pineda 2019)

2.2.1. Exemple de Dennett

En el primer cas, Dennett parla de diferents robots que tenen per objectiu aconseguir la bateria que els permet seguir funcionant, però tots ells acaben fracassant per algun motiu. El primer robot, R1, està programat per considerar la informació del seu entorn. Aquest descobreix que la bateria està dins d'un carro i que sobre el carro hi ha una bomba. Decideix arrossegar el carro i treure'l de l'habitació per agafar la bateria, però fracassa perquè no se n'adona que en arrossegar el carro, també està arrossegant la bomba. Aquesta acaba explotant i destrueix el robot. El segon robot, R1D1, és programat per considerar la informació disponible i les implicacions de les seves accions. R1D1 acaba fallant perquè es passa molta estona considerant totes les implicacions de les seves accions: treure la bateria de l'habitació no canvia el color de les parets, ni el radi de les rodes del carro, ni el pes de la bateria, ni moltes altres coses que romanen iguals. El robot es queda molta estona considerant les implicacions de les seves accions fins que acaba explotant la bomba. El tercer robot, R2D1, és programat per tenir en compte només les implicacions rellevants. Aquest acaba fallant, ja que per tenir en compte només les implicacions rellevants, abans ha d'haver descartat totes les implicacions irrellevants. I per haver descartat totes les implicacions irrellevants, abans ha d'haver analitzat totes les implicacions possibles i haver-les catalogat com a irrellevants. El robot considera una llarga llista d'implicacions irrellevants (color de les parets, temperatura de la sala, distància entre les parets, etc.). Es queda molta estona descartant implicacions irrellevants fins que, finalment, acaba explotant la bomba.

El problema que assenyala Dennett és que mai aconseguirem crear un robot que sigui capaç de descartar la informació irrellevant sense haver-la de considerar prèviament de forma explícita. Un humà no hauria d'avaluar quin és el color de les parets per saber que es tracta d'informació irrellevant per complir l'objectiu del robot, això és quelcom que l'humà descarta de forma automàtica. En canvi, un robot que està programat per considerar la informació que és rellevant no és capaç de dictaminar que certa informació és irrellevant sense haver-la avaluat prèviament. Aquest fet té a veure, precisament, amb el que Dennett entén per problema del marc: la incapacitat de descartar la informació irrellevant en una situació sense haver-la de considerar prèviament. Amb el seu exemple, Dennett acaba suggerint que mai aconseguirem obtenir un robot R2D2 com el que apareix

a la *Guerra de les galàxies*, el qual és capaç d'evitar el problema del marc amb tanta o més agilitat que un humà.

Cal veure que el problema del marc no consisteix en un problema tècnic; en el sentit que el robot no ha sigut capaç de descartar la informació irrellevant prou ràpidament. Encara que el robot hagués estat prou ràpid, el problema del marc seguiria existint, ja que el simple fet d'haver de considerar la informació de forma explícita per poder-la catalogar com a irrellevant és el que constitueix el problema del marc. Hom podria pensar que si el robot fos prou ràpid com per descartar la informació irrellevant, sense que es veiés mai compromès l'objectiu a assolir en cadascuna de les seves accions, tindria un comportament que no podríem distingir del comportament intel·ligent humà i potser es podria pensar que els humans també considerem i descartem la informació rellevant de manera molt ràpida.

Tanmateix, la hipòtesi anterior sembla poc plausible si considerem la forma en la que els humans considerem la informació. En una situació ordinària, sabem que hi ha certa informació que és del tot irrellevant sense haver-la de considerar prèviament, a diferència del robot. Això restringeix el domini d'informació candidata a ser considerada i fa que la resolució d'una tasca ordinària sigui més àgil que la d'un robot que consideri tota la informació explícitament.

El problema del marc ens fa plantejar com funciona la ment humana a l'hora de considerar la informació i prendre decisions. La intuïció que se'n desprèn és que no és el cas que la ment humana funcioni considerant la rellevància d'un gran nombre d'enunciats informatius en una situació ordinària, a diferència del robot R2D1. En situacions ordinàries, no considerem tota la informació disponible de la mateixa manera per tal de sentenciar que certa informació és irrellevant, sinó que deixem de banda aquella informació irrellevant de forma automàtica. Segons Dennett, la dificultat d'especificar quina és la informació rellevant (o de saber quina és la irrellevant) en cada moment és precisament el que constitueix el problema del marc i també és allò que fa que la IA no pugui arribar mai al nivell de refinament de la intel·ligència humana.

2.2.2. Definició del problema del marc

Els experts en IA veuen el problema del marc com un problema lògic: el problema de saber quins axiomes s'han d'establir en un model per determinar els aspectes del món que

romanen iguals un cop s'hi ha esdevingut una acció. Dennett entén que el problema del marc no és només un problema lògic, sinó que també és un problema concernent al coneixement. Defineix el problema del marc d'una manera que va més enllà de la dels lògics: el problema del marc consisteix en la incapacitat de descartar la informació irrellevant amb la mateixa habilitat que ho fa la intel·ligència humana.

Considero que s'han de complir dues condicions per poder dir que existeix el problema del marc: (a) ha d'existir un agent capaç d'avaluar la rellevància o irrellevància de certa informació i (b) aquest agent no ha de tenir la capacitat de catalogar certa informació com a rellevant o irrellevant abans d'haver-la avaluat.

La condició (a) parla d'un agent que tant pot ser humà com artificial (un sistema, robot, etc.). Independentment de quin sigui l'origen d'aquest agent (si és producte de la naturalesa o de l'enginyeria humana), l'important és que sigui un ésser capaç d'avaluar informació. Amb la condició (a) no en fem prou per dir que un agent pateix el problema del marc: si un agent descartés automàticament la informació irrellevant sense haver-la de considerar, aleshores no podríem dir que pateix el problema del marc. La condició (b) diu que l'agent no pot dictaminar el grau de rellevància de certa informació sense haver-la avaluat anteriorment. Cal veure que només amb la condició (b) no n'hi ha prou per dir que un agent pateix el problema del marc: una pedra compleix la condició (b), però no consideraríem que pateix el problema del marc. La conjunció de les condicions (a) i (b) ens diu que, perquè existeixi el problema del marc, cal que un agent amb capacitat d'avaluar el nivell de rellevància de certa informació només la pugui catalogar com a rellevant o irrellevant després d'haver-la avaluat.

Però el problema del marc és descrit per Dennett com el problema de filtrar la informació rellevant de la irrellevant amb la mateixa desimboltura que un humà. En una situació ordinària, un humà passaria per alt molta informació, ja que la consideraria dispensable. Un robot no té aquesta capacitat, ja que per poder donar compte de la irrellevància de certa informació, abans l'ha d'haver considerada (pensem en el cas del robot R2D1). Típicament, el problema del marc és detectat en sistemes informàtics per ressaltar com aquests han de considerar certa informació que els humans descartaríem automàticament (Pylyshyn 1987), però també pot existir el problema del marc en els humans si considerem casos més complexos.

El problema del robot que ha de treure la bateria de la sala és que el sistema no és capaç d'obviar la informació irrellevant, ja que no pot catalogar-la com a irrellevant sense haver-la examinat prèviament. Una altra qüestió seria preguntar-se per quins són aquests casos ordinaris: són casos ordinaris perquè no hi apareix el problema del marc o no hi apareix el problema del marc perquè són casos ordinaris?[1]

Generalment, el problema del marc no apareix en els humans. La ment humana té la capacitat de filtrar la informació rellevant de la irrellevant en casos ordinaris. Per exemple, quan hom s'adona que du les sabates descordades, és capaç d'avaluar la informació rellevant en aquella situació i obviar la irrellevant sense haver-la d'avaluar explícitament: és rellevant que m'aturi per cordar-me les sabates i poder seguir caminant, però és irrellevant saber qui és el president de Japó. La capacitat de filtrar la informació rellevant davant d'una contingència quotidiana sorgeix de forma automàtica en la majoria dels humans, així com la facultat de no considerar la irrellevant.

En la majoria de casos ordinaris, quan els humans no avaluem un enunciat és perquè l'enunciat conté informació irrellevant. En canvi, en el cas d'un robot programat per considerar els enunciats que versen sobre la informació disponible, la consideració d'un enunciat d'informació irrellevant només serà possible gràcies a la seva avaluació prèvia. L'explicació de per què la ment humana funciona així és un misteri que ha de resoldre la ciència cognitiva, potser una possible solució és que els humans hem après a considerar certa informació rellevant a través d'un procés evolutiu.

Hem pogut progressar evolutivament perquè hem sigut capaços de reaccionar de forma ràpida a situacions de vida o mort. La capacitat de reacció en moments de perill no hauria sigut tan àgil si hagués requerit descartar la informació irrellevant de forma explícita. Per exemple, un cavernícola que va sentir un soroll provinent de darrera els arbustos i va pensar que era una amenaça va poder sobreviure, en canvi, un company seu que va pensar que era un animal inofensiu va morir. A través de processos d'aquesta mena, s'ha anat determinant quina és la informació que cal considerar i quina és la que no s'ha de considerar. Igualment, un cavernícola que hagués estat considerant tot el ventall de

[1] Determinar què és ordinari és justament el que Hayes considera filosòficament rellevant del problema del marc, però això no és fonamental en el problema del marc (Hayes 1987). Fodor, en canvi, diu que el problema del marc és justament el problema de determinar quina és la informació banal (o la informació que apareix en qualsevol context ordinari) (Fodor 1987)

possibles candidats a ser els generadors del soroll que venia de darrera l'arbust, hauria mort si hi hagués hagut un animal perillós al darrera.

Tanmateix, els humans no estem exempts al cent per cent del problema del marc, hi ha moltes situacions en les que hem d'avaluar explícitament informació irrellevant per tal de catalogar-la com a irrellevant. Per exemple, imaginem que no recordem on hem guardat les claus del cotxe i comencem a pensar en el que hem fet durant l'última hora. Potser en aquest moment ens ve al cap què hem esmorzat. Aquesta informació és irrellevant en aquest cas, ja que no ens ajuda en res per acabar deduint que les claus són a la butxaca de la jaqueta. Tot i així, l'hem hagut de considerar per poder dir que era irrellevant, perquè estàvem avaluant el que havíem fet durant l'última hora i podria haver passat que en comptes d'esmorzar haguéssim canviat les claus de lloc. Per tant, en aquest context, era necessari avaluar aquesta informació per catalogar-la com a informació irrellevant; hem sabut a posteriori que era irrellevant, però només perquè l'hem considerada. En aquest treball, però, em limitaré a tractar situacions ordinàries com les del robot R2D1 en les que un humà descartaria implícitament aquella informació que el robot considera explícitament.

Dennett considera possibles candidats a ocupar el marc en el que es desenvolupen els nostres pensaments. Aquests candidats hauran de constituir coneixement i alguns d'ells seran més contextuals que d'altres. Per exemple, un coneixement que Dennett considera que ha de ser a-contextual és el de saber que si una cosa és a un lloc, no pot ser a un altre lloc simultàniament. Aquesta mena de coneixements són els que es troben en el fons de qualsevol raonament i poden ser considerats informació banal. Tanmateix, la IA força l'especificació de la informació banal, ja que el funcionament d'un sistema basat en IA s'ha d'especificar des de zero; els ordinadors que s'han de programar no coneixen res sobre el món, són una *tabula rasa*.

Un sistema basat en IA ben programat hauria de ser capaç de ser sensible al món i aprendre a reaccionar d'una manera determinada davant de cada situació. El dissenyador d'un sistema pot millorar el sistema tot afegint solucions per tal de tractar els casos particulars. És important notar que en aquests casos, el sistema no es redissenya a ell mateix o aprèn, sinó que és feina del dissenyador exterior; aquest procés de redissenyar s'assembla al procés de selecció natural en alguns aspectes, apunta Dennett. Fins ara, cap sistema d'IA ha sigut capaç d'aprendre autònomament a partir de la seva experiència passada:

> [...] and to date no one has been able to present any workable ideas about how a person's frame-making or script-writing machinery might be guided by its previous experience. (Dennett 1987)

La hipòtesi de Dennett és que la IA mai podrà arribar al nivell de la intel·ligència humana. Dennett creu que la ment humana no funciona de la mateixa manera que un sistema basat en IA programat per dur a terme certes tasques. Perquè puguem dir que la IA funciona de la mateixa manera que la ment humana cal que sigui possible adscriure-li una sèrie de propietats que considerem convencionalment idiosincràtiques al fenomen de la intel·ligència humana, una d'elles és la capacitat de resoldre el problema del marc en casos ordinaris. Segons Dennett, cap sistema és capaç de resoldre el problema del marc. Per tant, cap sistema basat en IA podrà ser assimilable a la intel·ligència humana; i.e., no podrem crear ments humanes a partir de la IA.

El meu objectiu és demostrar que la hipòtesi de Dennett és incompleta. Mostraré que Dennett té al cap un sol tipus de computació (computació convencional), mentre que obvia la computació evolutiva. Si existeix una forma de computar que no es veu afectada pel problema del marc, llavors no es pot descartar la possibilitat de crear una ment humana a través de la IA apel·lant a aquest problema. Argumentaré que la computació evolutiva no es veu afectada pel problema del marc i, per tant, no es pot descartar la possibilitat de crear un robot intel·ligent tenint en compte el problema del marc.

3. Tipus de computació i AE's

Podem distingir dues posicions filosòfiques envers la computació. Els que responen que cap forma de computació és filosòficament rellevant són els anomenats defensors de la *posició deflacionista* sobre les implicacions filosòfiques que es poden derivar de la ciència de la computació. En canvi, els defensors de la *posició anti-deflacionista* consideren que hi ha una forma de computació (computació evolutiva) que és filosòficament interessant, ja que planteja un cert estatus epistemològic privilegiat respecte la computació convencional (Muntean 2014). En aquest treball, assumiré que la posició anti-deflacionista és la correcta. Més endavant, explicaré què s'ha d'entendre per 'estatus epistemològic privilegiat', de moment em limito a apuntar que la computació evolutiva és un tipus de computació amb implicacions filosòfiques segons els defensors de la posició anti-deflacionista.

3.1 Computació convencional

La *computació convencional* (CC) consisteix a donar una sèrie d'ordres inicials (inputs) per tal d'aconseguir un resultat final (output). La CC és la mena de computació amb la qual estem més familiaritzats i és també la primera que ens ve al cap quan sentim a parlar de computació. Aquesta forma de computar permet conèixer des del primer moment l'output un cop s'han introduït els inputs necessaris. Dit d'altra manera, no hi ha res en l'output final que no estigui contingut en l'input. El resultat final de la programació es pot conèixer a priori per part d'un programador.

Tradicionalment, s'ha considerat que la CC no té cap interès per l'estudi filosòfic (Muntean 2014). La pregunta, llavors, és si existeix alguna forma de computació rellevant per l'estudi filosòfic. Per una banda, els qui adopten una posició deflacionista argumenten que la computació no mereix cap consideració filosòfica. La computació no té cap rellevància epistemològica per ells, ja que no és capaç d'oferir una resposta a la següent pregunta:

> P1: How can computers be made to do what needs to be done, without being told exactly how to do it? (Muntean 2014)

La pregunta P1 assenyala la incapacitat dels computadors de pensar per ells mateixos el que han de fer. Un computador no pot saber què ha de fer (en cas que poguéssim dir que té la capacitat de saber alguna cosa). Tampoc pot actuar com si sabés el que ha de fer sense que abans se l'hagi programat per tal cosa; no pot reflexionar sobre què convé fer en una situació. Darrere d'aquesta pregunta hi ha un pressupòsit:

> [1] A computer program is no better than the assumptions which it was built on. (Muntean 2014)

El pressupòsit [1] considera que tot el que conté un programa computacional ha estat especificat prèviament. Per tant, el programa actuarà només seguint els passos de la programació inicial i no podrà fer res pel qual no hagi estat programat. De manera que un sistema no pot ser millor, en el sentit de ser més intel·ligent, del que ha estat programat. Considero que el pressupòsit [1] és cert sempre que es té en ment la CC. Si un robot, per molts inputs que tingui per saber afrontar problemes, no té la capacitat d'aprendre per si mateix, llavors no podrem considerar que és intel·ligent:

> But the fulfillment of the logicist enterprise envisioned above was of just such a monster: one whose head is full of pre-ordained theories, driven by the universal logical crank of modus ponens, with no ability whatsoever to adapt to new surrounds. (Korb 1998)

A l'apartat 4, exposaré els motius pels quals la computació evolutiva nega [1] i, per tant, la pregunta P1 deixa de tenir sentit aplicada en el cas de la computació evolutiva.

Les màquines no pensen, no descobreixen i no inventen. Els deflacionistes constaten que els robots/màquines/sistemes estan basats en un conjunt de regles fixes en forma d'inputs que determinen els outputs resultants. Però les regles que determinen la producció de certs outputs a partir d'uns determinats inputs sempre són creades per un programador humà. Històricament, la forma més habitual de programar s'ha caracteritzat per ser determinista: no hi ha res en el resultat obtingut a través d'un procés de computació que no estigui contingut en el conjunt de regles preinscrites en el programa. Així, tota forma de programació és reduïble a un conjunt de regles (algoritmes) que donarà el mateix resultat independentment del moment i del computador en els que s'apliqui (Eiben & Smith 2003). En aquest sentit, el defensor de la posició deflacionista dirà que no hi ha cap element en el resultat d'un programa que escapi del coneixement del programador.

Per altra banda, els anti-deflacionistes asseveren que els defensors de la posició deflacionista tenen raó quan es refereixen a certs tipus de computació, però que s'equivoquen pel que fa la computació evolutiva. Ells fan una distinció entre dos tipus d'algoritmes: algoritmes com a proves formals i algoritmes com a processos de cerca d'optimització o d'aprenentatge. Els primers tipus d'algoritmes estan associats amb els algoritmes deterministes, típics de la CC. Els últims tipus d'algoritmes han inspirat l'àrea de la computació evolutiva. En aquest treball parlaré d'un tipus específic d'algoritmes basats en la computació evolutiva: els algoritmes evolutius.

3.2 Computació evolutiva

La *computació evolutiva* (CE) és una àrea de recerca dins la ciència de la computació que, tal i com el seu nom suggereix, està inspirada en el procés d'evolució natural. Aquest tipus de computació consisteix a simular un entorn omplert per una població d'individus que intenten sobreviure i reproduir-se[2]. En cada entorn, només sobreviuen els individus

[2] S'ha d'entendre *entorn* en el sentit d'entorn natural. En CE es pretén avaluar la competència dels individus exposant-los a un entorn per avaluar la seva capacitat de funcionament i de resolució d'una tasca

més competents. La competència de cada un dels individus està determinada per l'entorn i en com els individus s'acosten més o menys al seu objectiu, fet que determina la probabilitat d'un individu de sobreviure i passar a la següent generació.

La CE està inspirada en la teoria de l'evolució de Darwin, la qual descriu el mecanisme subjacent a la diversitat biològica: la selecció natural. Donat un entorn finit que no pot suportar més d'un determinat nombre d'individus, la selecció esdevé crucial per tal de preservar els individus més ben adaptats a l'entorn i eliminar els menys ben adaptats. La selecció natural afavoreix els individus que competeixen més efectivament per aconseguir els seus recursos. Un altre tret important per determinar la competència d'un individu, a part de la seva efectivitat, és el seu fenotip: la forma física d'un individu. La competència d'un individu determina la seva probabilitat de sobreviure i de reproduir-se a una generació ulterior. Darwin va apuntar que al llarg de l'evolució, s'esdevenen petites mutacions aleatòries en els trets fenotípics dels individus de cada espècie. Les millors mutacions sobreviuen, es reprodueixen, i, d'aquesta manera, evolucionen les espècies.

Es podrien comparar alguns elements de la CE amb la CC per veure algunes de les seves semblances i diferències. Per exemple, la CE es basa en la prova-i-error estocàstica, en canvi, la CC és fonamentalment determinista (no té cap component d'aleatorietat). La CC pot contenir alguna variable que es determini de forma aleatòria, però aquesta no serà essencial al seu funcionament. El que en CE anomenaríem l'entorn, l'individu i el grau de competència, en CC estaríem parlant del problema, el candidat a la solució del problema i el grau de qualitat, respectivament. Cal dir que parlar d'entorn, individu i competència són formes metafòriques per entendre intuïtivament el funcionament de la CE: es pot assimilar el genotip amb el codi de programació de l'algoritme i el fenotip amb el resultat obtingut. Així, per exemple, la solució a un problema matemàtic es podria representar fenotípicament amb el nombre 18 i genotípicament amb el codi 10010:

> A consequence of this view is that changes in the genetic material of a population can only arise from random variations and natural selection and definitely not from individual learning. It is important to understand that all variations (mutation and recombination) happen at the genotypic level, while selection is based on actual performance in a given environment, that is, at the phenotypic level. (Eiben & Smith 2003)

determinada, de manera semblant a la dels individus biològics quan són posats en entorns del medi i lluiten per sobreviure.

El mecanisme de selecció natural és molt eficient a l'hora de generar i millorar individus capaços de resoldre certs problemes. En aquesta línia, és comprensible que alguns científics de la computació s'hagin inspirat en un mecanisme com la selecció natural, capaç de millorar la qualitat de l'espècie mantenint els individus més ben adaptats:

> When looking for the most powerful natural problem solver, there are two rather obvious candidates: the human brain (that created the wheel, New York, wars and so on) and the evolutionary process (that created the human brain). (Eiben & Smith 2003)

3.3. Algoritmes evolutius

Per tal d'entendre el funcionament general dels AE's, dividiré aquest apartat en tres subapartats. El primer subapartat consistirà a definir de manera genèrica els AE's: parlaré del funcionament dels AE's i en comentaré les característiques principals. El segon consistirà en una revisió més detallada de la seva estructura tècnica. I, finalment, el tercer servirà per entendre el funcionament dels AE's a través d'un exemple.

3.3.1. Definició

Un *algoritme evolutiu* (AE) és una tècnica de recerca emprada en CE que té per objectiu trobar solucions completes o aproximades a problemes d'optimització.

> The combined application of variation and selection generally leads to improving fitness values in consecutive populations. It is easy to view this process as if evolution is optimising (or at least 'approximising') the fitness function, by approaching the optimal values closer and closer over time. (Eiben & Smith 2003)

Els AE's són uns tipus d'algoritmes basats en la CE, la qual fa servir tècniques inspirades en l'evolució biològica com la selecció, els creuaments o les mutacions. Hi ha molts tipus d'AE: algoritmes genètics (*GAs*), funcions reals basades en estratègies evolutives (*ESs*), arbres de programació genètica (*GP*), etc. Cada tipus d'AE tindrà un funcionament concret, però tots ells parteixen d'una base comuna que resumeixo en els paràgrafs que vénen a continuació.

Els AE's generen una població inicial aleatòriament. Avaluen tots els individus de la primera generació i seleccionen els més competents. Els individus seleccionats són creuats entre ells i/o mutats, el resultat d'aquests creuaments i mutacions es trasllada a la

següent generació. La resta d'individus (els que no han estat seleccionats) no es repliquen a la següent generació i són substituïts per altres individus generats aleatòriament. Aquest procés es repeteix fins que l'algoritme acaba. L'AE acaba quan s'ha complert cert nombre d'iteracions o quan un individu en concret ha aconseguit satisfer certa condició.

Dins d'un conjunt d'individus, un individu *x* serà més competent que un individu *y* si i solament si l'individu *x* està més ben adaptat a l'entorn que l'individu *y*. Estar més o menys adaptat a l'entorn dependrà de l'objectiu que es pretengui assolir en cada AE. En alguns AE, estar més adaptat a l'entorn significarà fer una tasca amb menys temps que la resta d'individus; en d'altres, significarà recórrer una distància superior, etc. Els AE's reprodueixen els individus a la següent generació tot emprant el principi de supervivència del més fort: es reprodueixen els individus més competents. Els AE's tendeixen a trobar el resultat òptim de cada procés tot reproduint els individus més competents al llarg d'un cert nombre de generacions.

Cal considerar, però, que hi poden haver errors de selecció de l'AE. Pot passar que un AE seleccioni un individu que hagi sobreviscut bé a l'entorn (sigui competent), però ho hagi aconseguit no per la seva qualitat, sinó perquè ha tingut sort:

> In EC, parent selection is typically probabilistic. Thus, high-quality individuals have more chance of becoming parents than those with low quality. Nevertheless, low-quality individuals are often given a small, but positive chance; otherwise the whole search could become too greedy and the population could get stuck in a local optimum. (Eiben & Smith 2003)

Aquest tipus de fenòmens poden esdevenir-se en un entorn natural quan, per exemple, un animal molt dèbil té la sort durant la seva existència de no topar amb cap enemic que pugui acabar amb el seu llinatge. En llenguatge evolutiu, diríem que el seu fenotip és competent, ja que hem vist que ha sobreviscut, però el seu genotip no és competent. Un resultat competent qualsevol ha de ser-ho tant a nivell fenotípic com genotípic.

3.3.2. Estructura

A continuació, exposaré l'estructura tècnica genèrica d'un AE. Aquesta consta d'una sèrie de passos que s'han de complir per aconseguir un objectiu determinat. L'esquema de funcionament dels AE's es pot representar de la següent manera:

```
INICI
      GENERAR població amb candidats a solució aleatòria
      AVALUAR cada candidat
      REPETIR FINS QUE ( CONDICIÓ DE FINALITZACIÓ sigui satisfeta ) DO
            1 SELECCIONAR pares;
            2 CREUAR parts dels pares;
            3 MUTAR la descendència resultant;
            4 AVALUAR els nous candidats;
            5 SELECCIONAR individus per la següent generació;
      OD
FINAL
```

Taula 1: Esquema general d'un AE en pseudocodi.

L'esquema anterior està constituït per ordres de programació que impliquen diferents accions a dur a terme per part de l'algoritme.

- L'ordre *GENERAR* serveix per produir un nombre inicial d'individus amb unes característiques aleatòries (genotip, en el cas de la selecció natural). Dos individus poden presentar fenotips idèntics, però això no implicarà que tinguin el mateix genotip, necessàriament. Tanmateix, si dos individus tenen el mateix genotip, llavors segur que tindran el mateix fenotip.
- La instrucció *AVALUAR* considera els diferents individus generats i avalua les característiques de cada un.

A continuació, es repetiran una sèrie de processos que no s'aturaran fins que s'hagi complert una *CONDICIÓ DE FINALITZACIÓ* (introduïda pel programador). En alguns casos, el procés s'aturarà quan s'arribi a un resultat òptim. Mentre que en altres casos, quan s'hagin complert certes condicions o s'hagi arribat a un nombre determinat de generacions. Pot passar que, atès que l'AE és estocàstic, mai s'arribi a un òptim. Aleshores l'algoritme programat per arribar a un òptim concret no s'aturarà mai. En aquest cas, s'haurà d'especificar una condició de finalització en forma de disjunció: l'AE s'aturarà quan arribi a l'òptim o quan s'hagin complert certes condicions.

1. El primer pas dels processos que s'aniran repetint és el de la selecció, exemplificat per l'ordre *SELECCIÓ*. Aquest operador és probabilista, i.e., opera

escollint els individus més ben adaptats segons certs criteris prefixats que variaran segons cada tipus d'AE amb una alta probabilitat, però també selecciona individus no tan ben adaptats. Això permet que es puguin tenir en compte característiques bones d'individus no tan ben adaptats. Funciona a nivell poblacional: seleccionarà els individus més competents de cada població perquè es reprodueixin a la següent generació i descartarà els individus menys competents, els quals seran substituïts per altres individus en la següent generació. Un individu serà pare d'algun membre de la següent generació si ha sigut seleccionat; la següent generació estarà constituïda per descendents de la generació anterior i per individus generats aleatòriament (la proporció de cada un d'ells dependrà de cada AE concret).

2. El segon pas, *CREUAR*, creua aleatòriament el codi de programació dels individus que seran els progenitors dels següents individus. És un operador estocàstic. Per exemple, l'individu progenitor primer podria tenir la codificació 10000100 i l'individu progenitor segon podria tenir la codificació 11111111. Aquests dos individus, en cas d'haver sigut escollits per l'operador de selecció, poden ser creuats a partir del tercer dígit, de manera que els dos individus resultants tindrien una codificació creuada de 10011111 i 11100100[3].

3. El tercer pas, *MUTAR*, canvia aleatòriament alguna de les parts de la codificació dels individus descendents. És un operador estocàstic, com l'anterior. Per exemple, l'individu descendent amb el codi 00000100 pot ser mutat si es canvia el seu segon dígit, de manera que s'obtingui l'individu amb la codificació 01000100[4].

4. La quarta ordre, *AVALUAR*, considera els individus que són descendència dels progenitors seleccionats, creuats i mutats, els quals són els individus de la nova població.

5. Finalment, la instrucció *SELECCIONAR* escull els descendents més competents de la nova població, que seran els progenitors de la següent generació. Aquest

[3] En biologia, la majoria dels organismes més complexos es reprodueixen sexualment. Això suggereix que la recombinació és una forma superior de reproducció. (Eiben & Smith 2003)

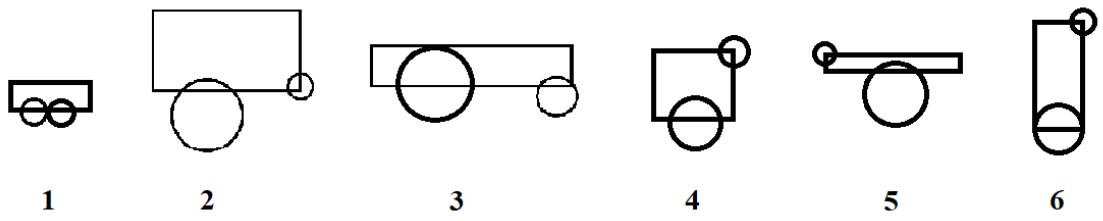
[4] During the recombination process, the choice of which pieces from the parents will be recombined is made at random. Similarly for mutation, the choice of which pieces will be changed within a candidate solution, and of the new pieces to replace them, is made randomly. (Eiben & Smith 2003)

procés arribarà al seu final quan s'hagi arribat a un nombre finit de generacions o bé s'hagi aconseguit solucionar un problema.

De l'esquema anterior podem deduir cinc característiques essencials dels AE's. (1) Els AE's operen a nivell poblacional, i.e., processen una col·lecció de candidats simultàniament. (2) La majoria dels AE's empren els creuaments, i.e., barregen la informació de dos o més solucions candidates per crear-ne una de nova. (3) Els AE's són estocàstics. (4) Encara que es parteixi d'una mateixa població inicial, els resultats als quals s'arribarà a través dels AE's no coincidiran en la majoria de casos degut a la aleatorietat del procés (tot i que res impedeix que passi). Les generacions inicials són aleatòries i independents de les generacions successives. (5) Hi ha una diferència important entre els AE's i el procés de selecció natural: mentre que en la selecció natural no hi ha cap direcció prefixada[5] –no actua seguint cap objectiu en concret, és cega–, no passa el mateix en el cas dels AE. Els AE's tenen un objectiu predeterminat pel programador; és a dir, no són completament estocàstics.

3.3.3. Exemple

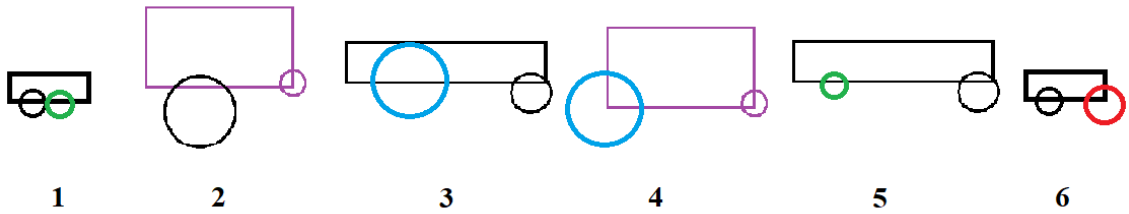
Suposem que un AE té l'objectiu de trobar un cotxe que aconsegueixi anar d'un punt A fins a un punt B. L'algoritme haurà de trobar un cotxe de dues rodes i un xassís rectangular que tingui les propietats idònies per poder recórrer el trajecte: forma del xassís, posició de les rodes i radi de les rodes. El trajecte és un camí amb pujades i baixades, de manera que hi haurà certes combinacions de propietats que donaran lloc a cotxes que quedaran encallats en el trajecte. Suposem que l'AE genera una població inicial de 6 cotxes i aquests són testats.



Il·lustració 1: Primera generació de l'AE.

[5] Segons la selecció natural i la teoria de l'evolució de les espècies no hi ha cap objectiu prefixat en la naturalesa. No entro a considerar teories que apuntin que la naturalesa segueix una direcció teleològica. Això és un tema d'estudi pertanyent a l'àmbit de la metafísica i la teologia.

L'AE seleccionarà aquells 3 cotxes que més metres hagin aconseguit fer sense quedar encallats. Suposem que els cotxes que han aconseguit recórrer més metres han sigut l'1, el 2 i el 3. L'AE reproduirà aquests 3 cotxes a la generació següent i els 3 restants seran creuaments i mutacions aleatoris de les diferents parts dels 3 seleccionats (el resultat es troba a la Il·lustració 2): el quart cotxe serà un creuament del segon cotxe (part lila) amb el tercer (part blava); el cinquè cotxe serà el creuament del primer (part verda) amb el tercer (part negra) i el sisè serà una mutació d'una roda del primer cotxe (part vermella). Les mutacions aleatòries possibiliten el sorgiment de solucions diferents a les que obtindríem amb només el creuament.



Il·lustració 2: Segona generació de l'AE.

Aquest procés s'anirà repetint fins que es compleixi la condició de finalització, la qual serà trobar un cotxe que arribi al final del trajecte. L'interessant d'aquest exemple, i dels AE's en general, és adonar-se que es pot arribar a solucionar l'objectiu de manera aleatòria. L'AE creuarà i mutarà aleatòriament els individus de cada generació. Per tant, els resultats de l'AE no tindran perquè coincidir en cada posada en marxa de l'algoritme.

També podem observar que la primera generació de cotxes en conté alguns amb poques probabilitats de recórrer gaire distància. Les formes dels xassís, així com les mides i posicions de les rodes dels cotxes 4, 5 i 6 són força estrambòtiques i s'assemblen poc a la forma d'un cotxe convencional. Tanmateix, no hem de descartar que hi pugui haver solucions òptimes poc versemblants al disseny d'un cotxe ordinari. Depenent de molts factors, com l'entorn, per exemple, obtindrem cotxes amb diferents formes. I ningú pot assegurar que el disseny humà de cotxes sigui més eficient que el que es pugui fer a través de la IA.

4. Els AE's i el problema del marc

A l'apartat 2, he explicat que el problema del marc sorgeix quan un agent, tot i tenir la capacitat d'avaluar informació, no té la capacitat de descartar la informació irrellevant sense haver-la d'avaluar prèviament. Això pressuposa que el sistema ha de tenir la capacitat d'avaluar algun tipus d'informació. Hom podria dir que un robot no pot avaluar la informació en el mateix sentit que ho fa un humà i que quan està fent el que és funcionalment idèntic a avaluar certa informació, en realitat, no l'està avaluant en absolut. Però si no pogués avaluar informació en absolut, llavors no patiria el problema del marc, ja que no compliria la condició (a), necessària per dir que existeix el problema del marc. En aquest treball, suposaré que el robot (o sistema) basat en IA té la capacitat d'avaluar la informació; és a dir, suposaré que la condició (a) es compleix pel cas del robot.

La CE té unes característiques que la diferencien de la CC. Crec que algunes d'aquestes característiques poden ajudar a resoldre el problema del marc. A continuació, exposaré els motius pels quals crec que els AE's poden complir aquesta tasca. El subapartat 4.1.1. representa una condició necessària per dir que els AE's poden aprendre, el subapartat 4.1.2. representa la superació del problema del marc i el subapartat 4.1.3. representa una similitud entre AE's i els humans.

4.1. Característiques rellevants dels AE's

4.1.1. Els AE's són autònoms

Els AE's són programats inicialment per un humà, però són els AE's els qui, de forma autònoma i aleatòria, intenten aïllar la solució a un problema concret. Aquesta és una característica compartida amb els humans: no hi ha ningú que controli les nostres accions de manera anàloga a com nosaltres controlem les accions de certes màquines. Aconseguir crear agents autònoms és l'objectiu principal de la IA:

> The ultimate goal of AI is to produce an autonomous articial agent which can cope with an a priori unknown world; hence, providing competent machine learning is a strict precondition for success. (Korb 2004)

La diferència és que en el cas dels AE's no hi ha cap humà al darrera que es preocupi d'anar modificant el codi de programació per tal de millorar la màquina en cada moment, sinó que és el propi algoritme l'encarregat de fer-ho. En aquest sentit, podem afirmar que els AE's són autònoms:

> In our view, there is even a more basic conception of autonomy which need not invoke the notion of goals or ownership, i.e., the notion of doing something for oneself. Instead, AUTONOMY need only be analyzed in terms of doing something by oneself, which pertains merely to independent behavior or activity, nothing more. (Muntean & Wright 2007)

L'AE fa la mateixa funció que faria un programador: revisar quina és la informació rellevant que permet solucionar els problemes davant d'una nova contingència. És la màquina l'encarregada de discernir entre la informació rellevant i la irrellevant. Justament perquè l'AE escull les millors màquines de cada generació és capaç de discriminar la informació rellevant de la no rellevant per tal de solucionar un problema concret.

Els AE's acaben produint un output que és considerat l'òptim dins d'un conjunt de possibles solucions. Aquest procés és aleatori, ja que es parteix d'una població d'individus que es van generant, creuant i mutant de forma aleatòria per donar lloc a un resultat. Imaginem que un AE ha d'optimitzar el disseny d'un cotxe que va des d'un punt A fins un punt B. De manera que el cotxe òptim serà el que recorri la distància que hi ha entre A i B en el menor temps possible i sense xocar amb cap obstacle. L'AE començarà amb una població inicial de 20 cotxes, per exemple, i reproduirà els 5 millors tot

combinant les seves parts (motor, codi de programació, xassís, etc.) per mirar d'optimitzar el resultat final. Imaginem, també, que aquest AE combina les parts del cotxe, posa els cotxes en marxa i els avalua mentre recorren la distància entra A i B. L'AE podrà combinar les rodes del primer cotxe més ràpid amb el motor del tercer, combinar el recorregut programat del segon amb el xassís del quart, etc. Així, els resultats tendiran cap a solucions (cotxes) que recorreran la distància en cada cop un període de temps inferior.

Ara, imaginem que durant el procés passa una persona aliena al procés d'optimització pel lloc on es fan les proves i és atropellada per un dels cotxes. Qui diríem que és moralment responsable de l'atropellament en un cas com aquest? Podríem dir que ha sigut l'AE, ja que el cotxe seria un producte del seu funcionament. Però també és cert que qui ha dissenyat inicialment l'AE ha sigut un humà. Semblaria que no està clar sobre qui recau la responsabilitat en aquest cas. Personalment, em decanto per la primera opció. Crec que només podríem imputar la responsabilitat a l'AE. A continuació, miraré d'argumentar per què crec que el vertader responsable moral de l'atropellament seria l'AE i no el programador.

Si partim de la premissa que la responsabilitat moral de l'atropellament recau sobre qui ha programat el cotxe, hem de dir que el culpable de l'atropellament és l'AE. El cotxe causant de l'atropellament és un dels cotxes particulars programats per l'algoritme de manera aleatòria. Podria haver sigut el cas que, en una altra ocasió amb les mateixes condicions de partida, l'AE hagués generat un cotxe totalment diferent que no hagués atropellat a ningú. Un cop el programador ha establert les condicions de partida de l'algoritme, l'AE és l'encarregat de trobar una solució a través del mecanisme de prova i error. L'AE genera una població inicial d'individus candidats a solució, els selecciona amb un criteri prefixat i en genera de nous, muta o creua aleatòriament. El component d'aleatorietat de l'AE fa que el resultat sigui completament independent del que el programador hagi pogut dissenyar prèviament. Si el cotxe particular que ha causat l'atropellament ha sigut generat per l'algoritme i el responsable moral de l'accident ha de ser el dissenyador del cotxe, llavors podem afirmar que el responsable de l'accident és l'AE particular que ha dissenyat el cotxe. Només podríem dir que el responsable moral de l'accident és el programador si aquest hagués programat l'AE amb la intenció d'atropellar algú.

Considerant l'aspecte de responsabilitat moral que podem reconèixer als AE's, no podem concloure que aquests algoritmes no pateixen el problema del marc. Però sí que podem considerar l'AE com un ésser autònom al qual se li poden atribuir propietats semblants a les que atribuiríem a un subjecte de decisió. Aquest punt és important per assenyalar el caràcter autònom que tenen aquests tipus d'algoritmes i que els distingeixen de la resta d'algoritmes convencionals. És important destacar que els AE's són autònoms justament pel seu component d'aleatorietat. Atès que el resultat de l'AE és majoritàriament aleatori, no podem dir que hagi estat dissenyat per ningú conscientment, tan sols podem dir que és un producte del propi AE.

4.1.2. Els AE's aprenen

Els AE's són uns algoritmes que optimitzen la solució a un problema. Per optimitzar la solució al problema, cal que adquireixin informació que sigui rellevant de cares al seu objectiu. Argumentaré, fent servir l'exemple de Dennett, que els AE's poden aprendre perquè són capaços d'adquirir autònomament informació rellevant.

Seguint l'exemple de Dennett, imaginem que hi ha una població inicial de 100 robots semblants a R2D1 en la majoria d'aspectes, menys en el fet que cada un d'ells considera un nombre finit d'informació de forma aleatòria. Hi ha un objectiu prefixat: aconseguir treure la bateria de l'habitació. Quan s'hagi aconseguit aïllar un robot amb el codi de programació adequat per complir l'objectiu, el procés finalitzarà. Pot passar que no s'aconsegueixi un robot capaç de complir aquest objectiu fins l'enèsima generació. Fins i tot, pot ser que mai s'arribi a aconseguir tal robot. Però, almenys, el que és segur és que un AE ben dissenyat optimitzarà el robot R2D1 apropant-lo a l'objectiu. No em centraré en els aspectes tècnics de com s'hauria de dissenyar l'algoritme, sinó que apuntaré el mètode general que seguiria un AE dissenyat per optimitzar la solució al problema:

En primer lloc, es generarà una població inicial de 100 robots capaços de considerar diversa informació aleatòria del seu entorn. En segon lloc, es reproduiran a la segona generació aquells robots que s'hagin acostat més al seu objectiu inicial. Com que l'objectiu inicial era treure la bateria de l'habitació, l'algoritme haurà reproduït el codi de programació d'aquells robots que s'hagin acostat més al seu objectiu: treure la bateria de la sala. Així, un criteri podria ser reproduir aquells robots que s'hagin acostat més a la porta de sortida amb la bateria. En tercer lloc, l'AE farà mutacions i creuaments entre els

robots seleccionats i en generarà de nous aleatòriament que substituiran aquells que no han sigut seleccionats. Aquest procés es repetirà fins aconseguir un robot que sigui capaç de complir l'objectiu o bé fins aconseguir un robot que s'apropi més a l'objectiu que la resta de robots.

Per aconseguir crear robots amb la capacitat de complir els objectius anteriors, caldrà que almenys algun d'ells hagi sigut capaç de considerar la informació rellevant i obviar la irrellevant. És a dir, hauran substituït informació rellevant per informació irrellevant i això els haurà apropat a l'objectiu. En aquest treball, defineixo *aprendre* com la capacitat d'adquirir coneixement proposicional. El coneixement proposicional serà útil si és informació rellevant. Així, el robot escollit per l'AE aprèn a complir l'objectiu (o s'hi acosta) perquè acaba adquirint la informació rellevant suficient.

> If a computer system is able -using a certain simulation of the scientific method- to present hypothesis that scientists find relevant, then the philosophical proposal embedded in the software has epistemic value. (Casacuberta & Vallverdú 2019)

Per la mateixa raó, diríem que hom aprèn a solucionar un problema concret: perquè reconeix la informació rellevant que el porta a la solució. La informació rellevant suficient per complir l'objectiu pot no ser necessària, ja que hi podria haver conjunts d'informació diferents que fossin suficients per complir l'objectiu.

L'AE aconsegueix superar el problema del marc, ja que aconsegueix captar la informació rellevant i descartar la informació irrellevant sense haver-la de considerar explícitament. Per una banda, dir que aconsegueix captar informació rellevant és equivalent a dir que aprèn. Això passa perquè l'AE reprodueix aquells individus que han sigut seleccionats per ser pares de les següents generacions. Els individus seleccionats ho seran en virtut de complir els criteris de selecció (determinats arbitràriament per un programador inicial). Si l'AE està ben programat, és a dir, si els criteris de selecció faciliten l'obtenció d'un individu que compleixi l'objectiu inicial, aleshores l'AE anirà aprenent a mesura que passin les generacions. És en aquest sentit que podem dir que els AE tenen un estatus epistèmic privilegiat respecte la CC:

> And one feature that [...] has been taken to most prominently differentiate the intelligent from the unintelligent is the ability to learn, to adapt behavior to new and challenging environments. If a robot, placed in an unanticipated environment, is unable to learn the

> simplest things about that environment and adjust to its simplest requirements, then that robot is not intelligent. (Korb 1998)

Per altra banda, l'AE aconsegueix descartar la informació irrellevant sense considerar-la explícitament, ja que selecciona aquells individus més competents. L'AE és capaç d'escollir els individus més competents un cop aquests han interactuat amb l'entorn. Per tant, l'AE no ha d'avaluar el grau de rellevància de cap conjunt d'informació, simplement es limita a seleccionar els individus més competents d'acord amb certs criteris de selecció.

És cert que algun dels robots anteriors hauria pogut complir el seu objectiu no per haver après a substituir la informació rellevant per la irrellevant, sinó que ho hauria pogut fer per pura sort. Tanmateix, hi ha la possibilitat de descartar aquests "falsos positius" tot exposant-los a noves situacions semblants a les anteriors. Per exemple, per saber que el robot que aconsegueix treure la bateria de l'habitació abans que exploti la bomba ha après realment a treure la bateria de l'habitació, es podria posar a prova l'habilitat d'aquest robot en la mateixa habitació (o en d'altres) amb modificacions d'objectes que catalogaríem com a informació irrellevant: el color de les parets, variant la temperatura, la intensitat de la llum, etc. Per tant, existeix un mecanisme, encara que resulti complicat, per corroborar la competència del robot escollit.

El problema dels "falsos positius" ens podria fer pensar en el problema de seguiment de regles (Kripke 1982), ja que en ambdós casos hi ha un agent (humà o robot) que actua com si sabés fer quelcom, però no és possible per un observador determinar si ho sap fer realment. Tot i no ser el mateix problema, comparteixen una mateixa dificultat: no poder verificar que algú sap sumar (cas del seguiment de regles) o que el robot sap fer quelcom sense avaluar infinits casos.

Sempre es poden imaginar noves contingències que acabin afectant a una màquina per molt competent que sigui, de manera que podríem arribar a pensar que el problema del marc segueix present. No obstant, gràcies als AE's, es podrà perfeccionar la màquina per solucionar les diverses contingències que la puguin afectar. Evidentment, no es podrà arribar mai a programar una màquina perfecta, però sí que es podrà solucionar un gran nombre de contingències a través del mecanisme de prova i error.

4.1.3. Els AE's generen resultats amb semblances rellevants

Una de les capacitats que intervenen en els processos d'aprenentatge és la d'establir semblances. No hem d'entendre la relació de semblança com una relació diàdica entre dos particulars (una relació absoluta), sinó que l'hem d'entendre com una relació de semblança relativa a un aspecte concret. Per tal d'establir la pertinença d'un particular a un conjunt de la mateixa classe, ens hem de basar en relacions de semblança relatives a l'aspecte del particular que pretenem incloure. L'espècie que té la capacitat d'establir semblances d'una forma més desenvolupada és l'espècie humana.

Podem pensar que els humans distingim amb més precisió certs objectes de la realitat d'acord amb una relació de semblança que la resta d'animals. Imaginem-nos un escenari en el que caiguessin meteorits capaços de destruir a l'instant tot allò amb el que impactessin. Per qualssevol dels animals no humans, els objectes que caurien a la Terra no serien més que perills que caldria esquivar per preservar la vida. Des del punt de vista d'un animal, no hi hauria diferència entre un meteorit amb un percentatge de 49% de ferro i un meteorit amb un percentatge de 50%. En canvi, els humans podríem establir classes de semblança pel que fa el percentatge de ferro del meteorit; imaginem que els meteorits de classe A serien els meteorits amb un percentatge de ferro de fins el 49% i que els meteorits de classe B serien els meteorits amb un percentatge de ferro estrictament superior al 49%. El que cal destacar d'aquest exemple és que els humans tenim l'habilitat d'establir relacions de semblança que poden ser útils per certes formes d'investigació (en aquest cas, científica) però que no són imprescindibles per la nostra supervivència. I és justament aquesta capacitat d'establir semblances la que ens permet avançar en els processos d'aprenentatge que van més enllà de l'aprenentatge necessari per la supervivència individual.

Aquest fet sembla allunyar la posició dels humans respecte la que ocupa la IA. Un algoritme informàtic convencional només operarà a través dels inputs que se li han assignat però mai no en podrà crear de nous. L'algoritme no podrà crear un patró de mesurament arbitrari com els que podem crear els humans. La manca d'aquesta capacitat arbitrària de creació de classes de semblança (que és indispensable per la majoria del coneixement humà) allunya la IA de la intel·ligència humana. Tanmateix, sembla que els AE's tenen la capacitat d'establir relacions de semblança de forma independent.

A través de l'operador *SELECCIÓ,* els AE's tenen la capacitat de seleccionar aquells individus que més s'ajustin a un paràmetre determinat. Suposem que en un primer moment existeixen cinc individus amb propietats diferents: x_1 = {A, D}, x_2 = {B, C}, x_3 = {A, C}, x_4 = {C} i x_5 = {A, B, C}, on A, B i C són tipus de propietats diferents. Imaginem que la millor propietat per complir l'objectiu de l'AE és la propietat B. Molt probablement, al cap d'unes quantes generacions, l'AE acabarà el seu procés convergint en una població on la majoria d'individus tindran la propietat B i l'individu òptim serà un individu amb la propietat B. En aquest cas, l'AE hauria sigut capaç de detectar una semblança rellevant que hauria de ser compartida pels diferents individus per poder ser considerats candidats a constituir un resultat òptim. Els individus més competents s'assemblarien pel fet de compartir la propietat B.

Els AE's tenen la capacitat de detectar semblances justament perquè l'operador *SELECCIÓ* aconsegueix fer convergir els candidats més competents. Els candidats més competents ho seran en virtut de tenir les millors propietats (aquelles que s'atansin més a l'objectiu) i s'assemblaran justament perquè compartiran aquestes propietats. Els AE's són capaços de detectar les semblances dels millors individus de cada població perquè distingeixen la informació rellevant de la irrellevant. És a dir, tenen en compte el tipus de semblança que compta per l'èxit de l'AE. De la mateixa manera que els humans, les màquines programades amb un AE saben en què s'han de fixar davant d'una contingència en concret. Això és possible perquè convergeixen a un resultat que tendeix a l'òptim i que, per tant, soluciona la contingència en la majoria de casos.

Però l'AE també podria establir infinites relacions de semblança de forma totalment irrellevant (sense cap implicació pràctica per la seva supervivència), tal com podem fer els humans. Podria donar-se el cas que l'AE seleccionés una propietat que no fos necessària per solucionar el problema (per exemple, C) només pel fet que aquesta ha anat acompanyada des de la primera generació amb la propietat òptima (B). En aquest cas, els candidats a la solució òptima compartirien una propietat no essencial per la seva supervivència; l'AE hauria detectat una semblança irrellevant per la supervivència, tal com fan els humans.

Els AE's generen resultats que, majoritàriament, s'assemblen en algun aspecte concret (solen compartir una propietat), ja que aquest aspecte és rellevant pel compliment de l'objectiu. L'AE té la capacitat de generar classes de semblança de forma autònoma, ja que no ha estat programat prèviament per seleccionar certes propietats particulars, sinó

que ha estat programat per seleccionar els individus més competents i reproduir-los. És a dir, ha estat programat per reproduir les propietats que cauen sota la descripció de ser millors propietats; les millors propietats en sentit *de dicto* (Quine 1956).

La capacitat de generar semblances de forma autònoma acosta els AE's als humans i, a la vegada, permet descartar el problema del marc. Si l'AE pot establir semblances que majoritàriament seran rellevants de forma autònoma, llavors té la capacitat d'arribar a un conjunt de candidats a solució òptima sense establir semblances irrellevants. No establir semblances irrellevants, implica no tenir en compte aquells individus menys competents; tenir en compte aquells individus amb informació irrellevant. Això ve a dir que l'AE no patirà el problema del marc perquè no considerarà explícitament la informació irrellevant per poder-la catalogar com a irrellevant, ho farà de forma automàtica i autònoma arribant a una solució amb individus que s'assemblen en alguna (o algunes) de les seves propietats.

4.2. Els AE's com a solució al problema del marc

L'objectiu general de la IA era crear una intel·ligència artificial real. Aquest objectiu, però, no es podia complir perquè existia el problema del marc. No es podia atribuir intel·ligència al robot R2D1 perquè no era capaç d'obviar la informació irrellevant amb la mateixa agilitat que un humà. Pels motius anteriors, semblaria que els AE's poden superar el problema del marc: els AE's permetrien que un robot aprengués adquirint el coneixement rellevant per complir una tasca sense necessitat d'una intervenció humana i obviaria la informació irrellevant sense haver-la de considerar explícitament.

Gràcies al component d'aleatorietat dels AE's, i a diferència de la CC determinista, la programació de R2D2 (el robot de Dennett programat a través d'un AE) respondria la pregunta P1 de l'apartat 3.1. i negaria el pressupòsit 1 que incorpora: no seria cert que tots els outputs de l'AE programats per dissenyar R2D2 es poguessin derivar dels inputs introduïts pel programador. El component d'aleatorietat de l'AE existeix gràcies a la generació aleatòria inicial d'individus i als operadors *CREUAR* i *MUTAR*. Aquestes característiques de l'AE fan que els resultats siguin impredictibles i no necessàriament coincidents entre diferents posades en marxa de l'AE. Un AE és programat per trobar la solució a un problema o per optimitzar un resultat, però no és programat per fer-ho d'una forma determinada.

Per tant, el problema del marc no constituiria un impediment per complir l'objectiu general de la IA:

> [...] although the frame problem is not yet solved in practice, it does not appear to offer any theoretical barrier to this approach to generating an artificial intelligence. (Korb 1998)

Solucionar el problema del marc és condició necessària però no suficient per aconseguir l'objectiu general de la IA:

> Hayes points out that the general AI problem – "GAIP" – is not the frame problem: that problem is the all-embracing one of how to build a real artificial intelligence. It is very clear that without solving the frame problem we will not solve GAIP, yet the solution of the frame problem is a specific precondition to a solution of GAIP, and not the very same thing. (Korb 1998)

Ara podem entendre per què Muntean (2014) atorgava un estatus epistèmic privilegiat als AE's respecte la resta de CC. Els AE's són capaços d'aprendre autònomament basant-se en un procés inspirat en la selecció natural. Aquests algoritmes poden arribar a descobrir per si mateixos lleis de la naturalesa o teoremes matemàtics ja demostrats científicament pels humans (Schmidt & Lipson 2009) i, qui sap, si podran descobrir-ne de nous abans que ho facin els humans.

El procés que segueixen els AE's per aprendre encaixa molt bé amb la hipòtesi que els humans hem après a descartar automàticament la informació irrellevant de manera evolutiva. Encara que, ara per ara, no es pugui contrastar la veracitat d'aquesta hipòtesi, reconec que és bastant plausible pensar que els éssers humans hem adquirit la capacitat de superar el problema del marc que pateixen robots com R2D1 a través de la reproducció d'aquells que no consideraven explícitament informacions irrellevants.

5. Conclusions

Crec que he aconseguit complir l'objectiu del treball, ja que he pogut argumentar que els AE's representen una forma de solucionar el problema del marc en IA. Els AE's ens brinden l'oportunitat de crear un sistema que no pateixi el problema del marc perquè són capaços d'aprendre de forma autònoma i descartar la informació irrellevant sense considerar-la explícitament. Emprant els AE's es podria aconseguir el robot R2D2 que deixava entreveure Dennett: un robot capaç d'aprendre a través del mètode prova-error i millorar el seu funcionament sense la necessitat d'intervenció humana.

L'objectiu de la IA, encara que segons la meva tesi no es pogués veure frustrat pel problema del marc, sí que podria descartar-se per altres motius que separen la IA de la ment humana: és pertinent atorgar estats intencionals a un algoritme?; es pot dir que la ment funciona a través d'una mera manipulació formal de símbols (Searle 1980)?; té sentit parlar d'una ment desconnectada d'un cos?; es pot crear una intel·ligència que iguali o superi la del creador (Chalmers 2010)?

Hi ha molts casos en els quals no està clar si un sistema o una persona pateix el problema del marc. Tanmateix, això no ha d'impedir l'estudi de casos més ordinaris en els quals sí estigui clar que certs robots pateixen el problema del marc descrit per Dennett. Que hi hagi casos de vaguetat (casos en els quals no sabem si és pertinent o no dir que hi ha el problema del marc) no invalida l'estudi de casos més generals.

He estudiat una possible solució al problema del marc tenint en compte la descripció de Dennett sobre el problema del marc. Un futur estudi relacionat amb el meu treball podria fer-se analitzant si els AE's poden solucionar altres visions del problema del marc de filòsofs com, per exemple, la de Fodor.

6. Bibliografia

Chalmers, D. (2010). The singularity: A philosophical analysis, *Journal of Consciousness Studies*, 17, 9(10):765.

Dennett, D. (1984), Cognitive Wheels: The Frame Problem in Artificial Intelligence, a Hookway, *Minds, Machines and Evolution*, Cambridge: Cambridge University Press.

Eiben, E.A., Smith J.E. (2015). *Introduction to Evolutionary Computing*, Nova York: Springer.

Hayes, P. & McCarthy, J. (1969). *Some Philosophical Problems from the standpoint of Artificial Intelligence*, Califòrnia: Stanford University.

Kamermans, M. & Schmits, T. (2004). *The History of the Frame Problem,* Amsterdam: University of Amsterdam.

Korb, K. (2004). Introduction: Machine learning as philosophy of science, *Minds and Machines*, 14(4): 433-40.

─ (1998). The Frame Problem: An AI Fairy Tale. *Minds and Machines*, 8: 317-51.

Kripke, S. (1982). *Wittgenstein on Rules and Private Language*, Massachusetts: Harvard University Press.

Muntean, I. (2014). *Computation and Scientific Discovery? A Bio-inspired Approach*, Indiana: University of Notre Dame.

Muntean, I. & Wright (2007) Autonomous agency, AI, and allostasis A biomimetic perspective, *Pragmatics & Cognition,* 15(3): 485-513.

Pineda, D. (2012). *La mente humana*, Madrid: Ediciones Cátedra.

─ (2019). *Sobre las emociones*, Madrid: Ediciones Cátedra.

Pylyshyn, Z.W. (1987). *The robot's dilema: The frame problema in artificial intelligence*, New Jersey: Ablex Publishing Corporation.

Quine, W.V.O. (1956). Quantifiers and propositional attitudes, *Journal of Philosophy*, 53: 177-86.

Schmidt, M. and Lipson, H. (2009). Distilling Free-Form natural laws from experimental data, *Science*, 324(5923): 81-5.

Searle, J. (1980). Minds, brains, and programs, *The Behavioral and Brain Sciences*, 3: 417-57.

Stanford Encyclopedia of Philosophy. (Febrer 2004). The Frame Problem. Recuperat de https://plato.stanford.edu/entries/frame-problem/

Vallverdu, J., & Casacuberta, D. (2019) Computational Philosophy as Experimental Philosophy, *International Journal of Innovative Studies in Sociology and Humanities*, 4(10): 23-33.