UNIVERSITAT DE BARCELONA

Ètica de les tecnologies: Coordenades teòrico-pràctiques per a la robòtica social

Júlia Pareto Boada

UNIVERSITAT DE BARCELONA

Ètica de les tecnologies: Coordenades teòrico-pràctiques per a la robòtica social

Tesi doctoral de

Júlia Pareto Boada

Directora i Tutora:

Dra. Begoña Román Maestre

Co-directora:

Dra. Carme Torras Genís

Facultat de Filosofia

Programa de doctorat en Filosofia Contemporània i Estudis Clàssics

4 de desembre de 2023

“El camí del pensament no recorre grans extensions. No pretén anar molt enllà, sinó una mica endins. I, just per això, reflexiona sobre el mateix, repeteix la mateixa cançó i manté el mateix horitzó.”

JOSEP MARIA ESQUIROL

“[…] el momento impulsor de esa forma de vida que denominamos «moderna» es la idea, el anhelo y el deseo de poner el mundo a *disponibilidad*. La vivacidad, la conmoción y la verdadera experiencia, sin embargo, surgen del encuentro con lo *indisponible*.”

HARTMUT ROSA

“Mentre que un cert intel·lectualisme cerca la novetat, el cor vol, sobretot, repetir.”

JOSEP MARIA ESQUIROL

AGRAÏMENTS

M'agradaria deixar constància de la meva gratitud a totes aquelles persones que han contribuït a elaborar aquesta tesi.

Molt especialment, voldria agrair:

A la Dra. Begoña Román, el seu suport i confiança absoluta, la seva generositat intel·lectual, la seva apassionada interlocució i el seu compromís amb el pensament com a acte de cura pel sentit.

A la Dra. Carme Torras, la seva excepcional valentia, esforç i lideratge per a vertebrar el tan necessari diàleg entre l'enginyera i la filosofia, així com per la seva orientació i acompanyament al llarg de tot el procés.

Al Dr. Josep Maria Esquirol, la seva lúcida custodia del tempo que li és propi al pensament, contra la lògica de l'acceleració i quantificació imperant; i per haver vetllat per a crear món amb tots i cadascun dels seus gestos com a coordinador del Grup de Recerca Aporia.

Al Prof. Mark Coeckelbergh, el seu recolzament i complicitat intel·lectual, i, sobretot, la seva amistat i alegre companyia a partir del moment en què em va acollir càlidament al Grup de Recerca Philosophy of Media and Technology de la Universitat de Vienna.

Al Prof. Peter-Paul Verbeek, la seva hospitalitat i la seva generositat i atencions per a integrar-me en el fantàstic equip que ha constituït a la Universitat de Twente.

Així mateix, agraeixo a l'Institut de Robòtica i Informàtica Industrial, i molt especialment al seu Director Guillem Alenyà, el vot de confiança en la Filosofia; i a Víctor Vílchez, Antonio Andriella, Pablo Jiménez i Siro Martín el seu incondicional suport.

D'altra banda, voldria afegir unes gràcies molt sentides, per haver ajudat decisivament a la compleció d'aquesta investigació, a Pol Pareto Boada, Jordi Arcos Pumarola, Juan Manuel Gómez Andrés, Eva Vidal, Camil Ungureanu i Núria Vallès Peris.

RECONEIXEMENT

Aquesta investigació doctoral ha estat possible gràcies al finançament rebut de:

PRE2018-084286 Ayuda para contratos predoctorales para la formación de doctores, finançada pel MCIN/ AEI/10.13039/501100011033, sota el segell d'excel·lència científica "María de Maeztu" (MDM-2016-0656) concedit a l'IRI (CSIC-UPC), i per “ESF Investing in your future”.

Projecte CLOTHILDE (CLOTH manIpulation Learning from DEmonstrations), Advanced Grant de l'European Research Council (ERC) dins de l'European Union’s Horizon 2020 research and innovation programme (grant agreement No 741930).

ÍNDEX

RESUM

Aquesta investigació doctoral, configurada com a compendi de publicacions, respon a la necessitat d'adreçar, des de l'ètica, el desplegament de la robòtica social, en especial pel que fa a la seva branca assistencial. Tal necessitat s'explica per l'emergència d'aquest camp de la robòtica intel·ligent com a proveïdor d'eines per a contextos d'activitat professional de l'àmbit de la salut, i per les insuficiències de la reflexió ètica que l'acompanya, que deriven d'una mala articulació disciplinària de l'ètica en relació a la robòtica social.

Davant d'això, la present tesi es planteja dos grans objectius, els quals es despleguen al llarg dels dos capítols principals que la conformen.

El primer objectiu és establir un marc conceptual per a una correcta aproximació ètica a la robòtica social, a través d'una doble tasca. D'una banda, sobretot en el Capítol 2, s'assenten els fonaments disciplinaris per a l'abordatge ètic d'aquesta tecnologia, a partir d'una clarificació de tres qüestions clau: per què la tecnologia demana d'ètica; de quin tipus d'ètica es tracta; i quin és l'estatut de l'ètica de la tecnologia? D'altra banda, en el Capítol 3, es delimiten unes coordenades ètiques específiques per a la robòtica social assistencial, és a dir, per identificar i analitzar les qüestions normativament rellevants per al seu desplegament, d'acord amb l'estatut d'ètica aplicada que li és propi a l'ètica de la tecnologia.

El segon objectiu és reexaminar qüestions ètiques centrals de la robòtica social assistencial, entrant en diàleg amb la discussió acadèmica actual. En vistes a aquest fi, en el Capítol 3 s'analitza críticament l'estat de la reflexió i les deficiències de l'aproximació ètica predominant, que es caracteritza per un oblit de la dimensió política d'aquesta tecnologia. En resposta als dèficits, s'examina la robòtica social assistencial des d'una de les coordenades ètiques anteriorment definides, la llibertat, ampliant el terreny de consideració normativa més enllà de l'esfera d'interacció diàdica humà-robot.

Complementàriament, davant la necessitat d'una innovació docent en ètica per graus universitaris d'enginyeria, es proposa un pla docent per a una assignatura d'ètica de la tecnologia de 6 crèdits ECTS.

ABSTRACT

This doctoral research, configured as a compilation of academic publications, responds to the need to address the deployment of social robotics from the ethics perspective, especially regarding its assistive branch. This need is explained by the emergence of this subfield of intelligent robotics as a tool provider for healthcare professional activities and by the deficiencies of the ethical reflection accompanying it, which derive from an improper disciplinary articulation of ethics on social robotics.

Therefore, the current thesis obeys two primary purposes, addressed through this work's two main chapters.

The first goal is establishing a conceptual framework for a proper approach to social robotics through a double task. On the one hand, mainly in Chapter 2, the disciplinary foundations for the ethical approach to this technology are grounded through clarifying three key questions: why does technology require ethics, what kind of ethics is it, and what is the statute of technology ethics? On the other hand, in Chapter 3, some specific ethical coordinates for social assistive robotics are defined; that is, coordinates to identify and analyse the normatively relevant issues for its deployment, in line with the status of applied ethics appropriate to the ethics of technology.

The second goal is to reexamine some central ethical issues of social assistive robotics, thereby entering into dialogue with the current academic discussion on this theme. To that end, Chapter 3 puts forward a critical analysis of the reflection on this technology and the deficiencies of the prevailing ethical approach, which is characterized by a neglect of the political dimension of technology. In response to such deficits, social assistive robotics is examined from one of the previously defined ethical coordinates, namely freedom, thereby expanding the scope of normative consideration beyond the sphere of dyadic human-robot interaction.

Additionally, given the need for innovative teaching programs on ethics in university engineering degrees, the thesis proposes a teaching plan for a technology ethics subject of 6ECTS credits.

CAPÍTOL 1. INTRODUCCIÓ

1. La robòtica social com a nova interpel·lació a l'ètica

En el context contemporani d'una creixent capacitat de la robòtica com a proveïdora d'eines per a l'activitat humana, resulta especialment significativa l'emergència de la robòtica social. Si haguéssim de fer una lectura de la nostra circumstància tècnica en clau epocal semblaria, si més no, que la robòtica social representa un nou episodi del moment històric que alguns filòsofs contemporanis van designar, ja al s. XX, com l'era de la tècnica (Heidegger, 2021) (Ortega y Gasset, 2004). En efecte, el desenvolupament de robots capaços d'interactuar amb els humans de forma "interpersonal"(Breazeal, Takanishi and Kobayashi, 2008) per a servir en diferents contextos pràctics suposa un pas més cap a la configuració tècnica del món a què fa referència la caracterització filosòfica de la contemporaneïtat. Avancem, ara, cap a la tecnificació d'àmbits d'activitat humana que, per la seva naturalesa relacional, quedaven encara substantivament reservats a l'agència humana.

Aquesta darrera manifestació de la nostra condició tècnica suposa una interpel·lació a l'ètica actual. Contra l'acriticisme imperant amb què contemporàniament es pressuposa el sentit d'aquesta interpel·lació, aquesta no és, tanmateix, una afirmació evident. Clarificar-la ocuparà un lloc central en aquest treball. D'entrada, però, convé destacar que la interpel·lació es diu aquí en la significació més completa del terme: no es tracta només que a l'ètica, com a activitat de fonamentació critico-racional de la moral, se li demani orientació per a la presa de decisions a què ens confrontem amb les noves possibilitats obertes per la robòtica social. Es tracta, també i més fonamentalment, que l'ètica es veu sol·licitada a realitzar una tasca d'autoreflexió i aclariment disciplinar: es veu impel·lida a donar explicacions del seu estatut i activitat en relació a la robòtica social.

Per entendre la interpel·lació en tota la seva vastitud, resulta oportú començar per contextualitzar-ne la seva particularitat. Això demana, abans que res, situar-la en el marc d'una llarga relació entre tecnologia[1] i ètica.

1.1. Tecnologia i ètica: Una relació amb història

Si bé ha anat experimentant canvis significatius pel que fa al caràcter de la seva articulació, la relació entre tecnologia i ètica no és nova.

Les implicacions de la tecnologia per als individus i les societats han estat objecte d'atenció i reflexió en el pensament filosòfic ja des dels seus inicis. Recordem, per exemple, la meditació

[1] Davant la varietat d'accepcions a què respon la distinció terminològica entre tècnica i tecnologia, en aquesta tesi s'opta per usar, en general, el segon dels termes. Aquesta decisió s'explica, d'una banda, pel mateix objecte d'estudi de la tesi –la robòtica social (assistencial)–, que correspon a un tipus d'activitat i artefactes propis de la tècnica que apareix al s. XX (l'anomenada "tècnica moderna") i que es troba constitutivament vinculada a la ciència. D'altra banda, la decisió es pren estratègicament per mantenir una millor consonància amb la terminologia (anglosaxona) predominant en l'activitat filosòfica contemporània entorn la tecnologia –en què el terme referencial és "technology"– i amb la literatura acadèmica amb què aquesta tesi entra en discussió. Amb això, es facilita que la tesi se situï com a interlocutora amb la filosofia i l'ètica de la tecnologia actual. En aquesta tesi, el terme "tècnica" s'usarà exclusivament en aquells casos en què, per motius històrics i bibliogràfics, sigui més adient.

sobre la tècnica de l'escriptura i les seves possibles conseqüències negatives que Plató desenvolupa en el Fedre (Plató, 1988), un diàleg que data del 370aC. Recorrent al mite de Thamus i Theuth, i agafant de referència el cas de l'escriptura, la reflexió platònica subratlla l'ambivalència de la tècnica, és a dir, el fet que aquesta pot tenir resultats no només no desitjats ni desitjables, sinó inclús contradictoris amb l'objectiu per als quals es concep. Aquesta és una característica de la tècnica fonamental des d'un punt de vista ètic, en tant que, d'entrada, demana evitar adoptar un tecnooptimisme acrític –com el que representa la figura de l'inventor Theuth– i, conseqüentment, situa el desplegament i ús de la tecnologia dins el terreny de consideració normativa.

L'anàlisi platònic de l'escriptura, més enllà d'il·lustrar la historicitat de la reflexió filosòfica sobre la tecnologia, destaca per ser notablement representatiu del caràcter de gran part de la consideració ètica vers la tecnologia desenvolupada per l'activitat filosòfica. En efecte, és predominantment amb motiu dels potencials riscos que se'n deriven que la tecnologia es defineix com a tema ètic singular en la història de la filosofia.

En aquest sentit, no resulta estrany que sigui precisament al s. XX, davant la magnitud de la capacitat de la tècnica moderna[2] i de les seves possibles conseqüències –que esdevenen, en la majoria dels casos, imprevisibles i irreversibles, i fins i tot radicalment transcendents per la vida humana–, que es formuli, a mans de Hans Jonas, un replantejament de la disciplina de l'ètica. Les dimensions que adquireix l'acció humana a partir dels avenços tecnològics requereixen d'una nova ètica capaç d'orientar-la corresponentment. En una situació en què la nostra capacitat de fer sobrepassa la nostra capacitat de saber –això és, de preveure'n els efectes últims–, aquesta serà una ètica articulada entorn al principi de responsabilitat –l'anomenada "ètica de la responsabilitat" (Jonas, 2015, p.354).

Tornant a l'activitat de reflexió ètica sobre tecnologia pròpiament, malgrat remuntar-se a l'Antiguitat i perdurar en èpoques posteriors a mans d'autors com Francis Bacon i Jean-Jacques Rousseau[3] (Coeckelbergh, 2020), no és fins al s. XX que es desenvolupa l'ètica de la tecnologia com a branca específica de la disciplina filosòfica (Franssen, Lokhorst and van de Poel, 2023). Fins llavors, les consideracions ètiques sobre tecnologia s'havien desenvolupat com a part de l'activitat d'altres branques de la filosofia, en consonància amb el que havia passat amb la dedicació filosòfica general a l'entorn de la tecnologia (Mitcham, 1994, p. 17). Aquesta última no va emergir com a subdisciplina filosòfica fins al s. XIX, quan es formalitza oficialment com a Filosofia de la tecnologia (Coeckelbergh, 2020) en base a la terminologia encunyada per Ernst Kapp (Kapp, 1877). La sistematització del pensament filosòfic sobre tecnologia comença, per tant, en un context històric de grans canvis tecnològics i socials a partir de la Revolució industrial

[2] D'acord amb l'ús que se li dona dins el pensament filosòfic, l'expressió "tècnica moderna" fa referència al tipus de tècnica que apareix al s. XX (Jonas, 2015) (Heidegger, 2009), caracteritzada per comportar un canvi substantiu per a la nostra manera de configurar el món. En aquest sentit, no s'ha de confondre amb la connotació historiogràfica que té el terme 'modern' en el marc de la tradició filosòfica.

[3] La tecnologia rep una atenció considerable per part de la filosofia sobretot en el període modern, si bé des d'una perspectiva positiva que és, doncs, substantivament diferent a la que marcarà l'inici disciplinari de la filosofia de la tecnologia (Brey, 2010). Com s'aclarirà en el Capítol 3, aquesta aproximació positiva va vinculada a la comprensió moderna de la tecnologia com a axiològicament neutra (Feenberg, 2018), la qual explica, al seu torn, la tardana constitució de la tecnologia com a camp propi de l'ètica.

i s'acaba de consolidar i desenvolupar fortament al s. XX, marcat també per grans esdeveniments en què la tecnologia hi juga un rol fonamental.

La demarcació de la tecnologia com a camp específic de reflexió ètica coincideix, doncs, en bona mesura, amb el "gir aplicat" (Cortina, 2003) que va viure l'ètica al s. XX. Això ajuda a entendre certs trets de la configuració disciplinària de l'ètica de la tecnologia.

Per començar, explica que la tecnologia es defineixi com a objecte singular de l'ètica a partir del moment en què aquesta es replanteja a si mateixa com a prioritàriament pràctica, és a dir, destinada a orientar l'acció humana davant els problemes i complexitats de la nostra realitat quotidiana (Kettner, 2003). Atès el seu potencial disruptiu i la centralitat que adquireix en els diferents àmbits de la vida humana, la tecnologia passa a formar part dels problemes de l'ètica contemporània, és a dir, del conjunt de qüestions normativament rellevants que, com a tal, requereixen de la seva tasca d'orientació pràctica.

Per consegüent, s'entén també que l'ètica de la tecnologia es conceptualitzi predominantment com a branca de l'ètica aplicada (Franssen, Lokhorst and van de Poel, 2023) (Gordon and Nyholm, 2023) (Sætra and Danaher, 2022) i que, en sintonia amb el "gir empíric" de la filosofia de la tecnologia de finals del s.XX (Verbeek, 2010, p.49), procedeixi a aproximar-se a la tecnologia ja no com a fenomen en si –com a "Tecnologia", com feia la filosofia de la tecnologia clàssica (Brey, 2010, p.39)– sinó en la particularitat de cadascuna de les seves concrecions. Això explica la dualitat que s'identifica comunament en el domini de l'ètica de la tecnologia, que es diversifica en dos camps d'estudi (o d' "aplicació") principals: d'una banda, l'activitat professional de desenvolupament tecnològic (ètica de l'enginyeria) i, de l'altra, les tecnologies específiques. Aquestes donen lloc, al seu torn, a una multiplicitat de subdominis de l'ètica de la tecnologia (Verbeek, 2011), que es concreta així en ètica de la computació, ètica de la IA, ètica de la robòtica, ètica de la ciència de dades, ètica de la nanotecnologia, ètica de la biotecnologia, ètica de la geoenginyeria, etc.

Significativament, però, la relació entre tecnologia i ètica fa temps que no es defineix únicament per l'activitat reflexiva desenvolupada per la filosofia. La relació, en aquest sentit, no és unidireccional. Ja no es tracta només que les humanitats en general, i la filosofia entre elles, es preocupin per les implicacions del desplegament tecnològic –i, sobretot més recentment, per ajudar a orientar-lo adientment–. Contemporàniament, és també des de l'àmbit de l'enginyeria que es reivindica la necessitat de reflexió ètica i se l'integra dins el propi camp de treball professional[4].

El mateix naixement de la Roboètica (Veruggio, 2006) n'és un clar exemple. Aquesta disciplina, definida com una "ètica aplicada a la robòtica" (Veruggio, Solis and Van Der Loos, 2011), va ser concebuda i formulada l'any 2004 per part de la comunitat d'enginyers, en resposta al gran creixement del seu camp i les noves qüestions ètiques plantejades, sobretot, per la robòtica intel·ligent. Així doncs, l'articulació d'una reflexió ètica específica sobre la robòtica emergeix

[4] És oportú precisar que aquesta és una aproximació diferent i posterior a la de l'anomenada "engineering philosophy of technology" (Mitcham, 1994), una de les dues tradicions històriques de la filosofia de la tecnologia, caracteritzada per interpretar i explicar el món des d'un paradigma tecnològic.

primàriament de la preocupació del propi sector professional per a orientar-ne el desplegament en vistes al benefici de la humanitat (Operto and Veruggio, 2008)(Tzafestas, 2018).

Sense entrar en una anàlisi històrica exhaustiva de l'evolució del compromís de l'enginyeria amb una reflexió crítica sobre la pròpia activitat, resulta interessant assenyalar que el cas de la Roboètica és simptomàtic d'un canvi qualitatiu en aquest sentit. Més enllà dels passos per una ètica de la professió al s. XX –que, malgrat materialitzar-se de forma i a ritmes diferents en el continent americà i europeu (Didier and Heriard-Dubreuil, 2005), tenen a veure amb una trajectòria global de revisió i desenvolupament de codis ètics d'enginyeria i, posteriorment, d'introducció de cursos d'ètica en programes universitaris d'enginyeria (Mitcham and Briggle, 2009)–, la responsabilització de l'enginyeria per l'impacte social del desenvolupament tecnològic es tradueix, al s. XXI, en una integració particular de la reflexió ètica en la dinàmica professional dels enginyers. El canvi substantiu està en què, a banda de l'exercici d'una ètica de la responsabilitat individual i professional –fonamentada pels codis ètics i l'educació en capacitats crítiques a partir de cursos d'ètica a les universitats politècniques–, els enginyers tendeixen a desenvolupar la seva activitat en un context de col·laboració interdisciplinària, per a identificar i adreçar les qüestions normativament rellevants del desplegament tecnològic de manera més comprehensiva (Torras, 2024). Aquest és el plantejament amb què neix la Roboètica –que, si bé a la pràctica ha estat majoritàriament desenvolupada per enginyers, en la seva concepció es va formular com a tasca interdisciplinària (Veruggio, 2006) (Veruggio and Operto, 2006).

És en el marc d'aquesta relació entre tecnologia i ètica, doncs, que cal interpretar l'emergència de la robòtica social com a nova interpel·lació a l'ètica contemporània. D'entrada, queda clar que l'aparició d'aquesta branca de la robòtica intel·ligent interpel·la l'ètica *novament*. És a dir, suposa una interpel·lació més d'entre les que ja se li han formulat i les que vindran a continuació amb motiu del desenvolupament de diferents camps tecnològics i les seves implicacions per als diversos àmbits de la vida humana. Resta per determinar, ara, el sentit en què es pot considerar la interpel·lació com a *nova* en l'accepció més radical del terme, això és, pel fet de resultar d'una circumstància tecnològica substantivament diferent de les anteriors, almenys en cert grau. Això passa per clarificar la particularitat de la robòtica social com a tecnologia de potencial disruptiu inèdit que, en conseqüència, reclama d'una reflexió ètica específica.

1.2. El cas de la robòtica social

La robòtica social és una branca de la robòtica intel·ligent força recent que, no obstant, està en ple desenvolupament (Mejia and Kajikawa, 2017). Malgrat haver abandonat definitivament el podi de la ficció i tractar-se, ja, d'una realitat (emergent), la robòtica social s'enfronta a una certa ambigüitat pel que fa a la seva definició. Això es deu a la manca de consens en la caracterització de l'objecte primari de la seva activitat, els robots socials (Torras, 2024) (Sarrica, Brondi and Fortunati, 2020), així com, encara més fonamentalment, de la mateixa noció de robot (Bekey, 2014) (Coeckelbergh, 2022). No obstant, es poden assenyalar almenys tres aspectes vertebrals de la robòtica social, en base als quals aquesta es caracteritzaria com a camp tecnocientífic consistent en (1) desenvolupar sistemes d'intel·ligència artificial (IA) corporeïtzada, (2) capaços d'interactuar amb els humans de forma "interpersonal", és a dir, seguint els patrons del que seria una comunicació intersubjectiva significativa, i (3) destinats a tasques concretes en activitats humanes diverses.

Resulta important aturar-se a clarificar els matisos de cadascun d'aquests trets definitoris de la robòtica social.

En primer lloc, com a robots intel·ligents[5], els objectes de la robòtica social són agents artificials que poden realitzar conductes en entorns reals amb un cert grau d' "autonomia tecnològica" (European Parliament, 2017, p.242) –és a dir, sense necessitat de control extern–, com a resultat d'una prèvia percepció de l'entorn i d'un processament de la informació sobre aquest i sempre en vistes a un objectiu que els ve predeterminat externament. D'entrada, doncs, els robots socials són dispositius físics amb capacitat d'assumir tasques no només mecàniques, sinó que requereixen d'interacció i adaptabilitat a l'entorn.

Que els objectius als quals respon el funcionament dels robots estiguin ja predeterminats implica que el tipus d'intel·ligència artificial d'aquests robots és la denominada IA "estreta" (o "dèbil"). Es tracta de la capacitat de dur a terme tasques molt específiques en àmbits molt concrets, i és, per tant, totalment diferent del que es coneix com a IA "general" (o "forta") (High-Level Expert Group on Artificial Intelligence, 2019, p.5), que seria aquella capaç de realitzar qualsevol de les tasques cognitives que poden fer els éssers humans. Aquesta apreciació és rellevant per a una correcta priorització temàtica en la reflexió ètica sobre robòtica social, allunyada de la d'una activitat merament especulativa (van der Plas, Smits and Wehrmann, 2010) (Nordmann and Rip, 2009).

En segon lloc, en relació a la seva capacitat d'interacció social, és important subratllar que aquesta és, precisament, el mitjà a partir del qual els robots desenvolupen les seves tasques. D'aquí que s'anomenin robots "socials". Contràriament al que sovint es pressuposa, l'etiqueta "social" no respon primàriament al fet que aquests robots s'integrin en contextos de vida quotidians, a diferència dels industrials. Això portaria a prendre'ls per equivalents amb els robots de serveis, quan aquests poden, no obstant, desenvolupar les seves tasques sense necessitat d'interactuar amb els humans i, per tant, no tenen per què ser necessàriament "socials". Si es categoritzen com a "socials", doncs, és perquè, seguint la forma de comunicació humana, els robots es relacionen amb les persones per a realitzar tasques que requereixen, en major o menor grau, d'una interacció personal.

Això entronca essencialment amb el tercer aspecte de la definició, que, com s'anirà clarificant, és fonamental des d'un punt de vista ètic: en tant que la interacció és un mitjà per al desenvolupament d'una tasca, la interacció cau sempre dins un marc de funcionalitat-finalitat. És a dir, en última instància, la interacció social del robot es concep en vistes a servir a un objectiu determinat, en el marc d'activitats humanes diverses: proveir assistència davant necessitats especials (físiques o cognitives), fer companyia, ajudar en processos d'aprenentatge, satisfer necessitats sexuals, etc.

Els robots socials, doncs, poden dur a terme funcions diferents i aplicar-se a una gran varietat de contextos pràctics. Actualment, la robòtica social s'està desenvolupant sobretot com a proveïdora d'eines per a contextos d'activitat professional. En gran mesura, la recerca i desplegament d'aquesta branca de la robòtica respon a l'objectiu de contribuir tecnològicament

[5] Malgrat l'absència d'unanimitat en la seva definició, el paradigma predominant amb què es caracteritzen els robots (intel·ligents) per part del correlatiu camp de l'enginyeria és el del "sentir, pensar i actuar" (Bekey, 2014, p.18).

a pràctiques institucionals, especialment a aquelles vinculades a l'àmbit de la salut. A Europa, això es troba en sintonia amb l'interès declarat per desenvolupar la IA, la robòtica i la digitalització prioritàriament per a aquest sector (European Commission, 2020) (Dolic, Castro and Moarcas, 2019). L'aposta europea pel desplegament tecnològic en salut té a veure amb la creixent demanda de servei a què s'enfronta el sector sociosanitari, molt especialment degut a l'increment de població amb necessitats especials; un fenomen que, al seu torn, s'explica en gran mesura per l'envelliment de la població global (United Nations Department of Economic and Social Affairs, 2022) (Caleb-Solly, 2016) . En aquesta línia, una de les branques de la robòtica social en major desenvolupament és l'assistencial (Torras, 2019), i no és d'estranyar que sigui la cura de la gent gran un dels seus àmbits d'aplicació principals (Vandemeulebroucke, Casterle and Gastmans, 2020) (Caleb-Solly et al., 2014) .

En termes generals, la robòtica social assistencial (SAR)[6] s'orienta a facilitar robots que interactuen amb els humans com a mitjà per desenvolupar tasques que tenen a veure amb l'ajuda a persones amb necessitats especials. Dins aquest grup s'hi inclou gent gran amb cert grau de dependència, persones amb demència, pacients convalescents, persones amb patologies de salut mental, infants amb trastorn de l'espectre autista, i persones amb diversitat funcional. Així, aquests robots es conceben per ser introduïts en diferents activitats assistencials de l'àmbit de la salut –com la teràpia i la rehabilitació física i cognitiva, l'educació especial o suport a la dependència–, i en diferents ecosistemes de cura –hospitals, residències de gent gran, escoles o entorns domèstics–. En aquests contextos, els robots socials poden desenvolupar funcions i rols variats, com el d'entrenador en exercicis de restauració o manteniment de la salut, el d'assistent en tasques quotidianes com menjar o vestir-se, o el de company en tractaments de salut mental (Rabbitt, Kazdin and Scassellati, 2015). Alguns exemples d'aquest tipus de sistemes d'IA socialment interactius són el *Pepper*, el *Nao* i el *Paro* –emprat per a la teràpia amb gent gran i il·lustratiu de la varietat morfològica d'aquests robots, que no tenen per què ser necessàriament humanoides–.

Tot i que la seva implementació encara no és una realitat consolidada, a Europa existeixen diverses iniciatives i projectes pilot remarcables en robòtica social assistencial, que evidencien les expectatives institucionals d'introduir aquests artefactes en l'activitat assistencial de l'àmbit de la salut, especialment en relació a la cura de la gent gran.

N'és un exemple rellevant el sistema d'IA robòtic desenvolupat per l'Institut de Robòtica i Informàtica Industrial (CSIC-UPC), amb la col·laboració de la Fundació ACE de Barcelona, com a eina de suport a l'activitat terapèutica (Andriella, Torras and Alenyà, 2020). Concebut en el marc del projecte europeu SOCRATES, es tracta d'un robot social assistencial que interactua amb els pacients per a ajudar-los a realitzar un exercici d'estimulació cognitiva que forma part de la tasca assistencial de la Fundació ACE. Un altre cas il·lustratiu és el projecte pilot de l'Ajuntament de Barcelona per a millorar, mitjançant la implementació d'un robot social assistencial a la llar (Fundació iSocial, 2020), la qualitat de vida de la gent gran amb algun tipus de dependència que

[6] Procedents de l'anglès *social assistive robotics*, SAR són les sigles de referència per a designar aquest camp, i les que s'adopten, per tant, en aquesta tesi –sobretot per al seu ús en els articles acadèmics derivats–. Respectivament, SARs es prenen com les sigles dels seus artefactes, els robots socials assistencials (*social assistive robots*).

viu sola –una iniciativa que comença a desplegar-se pel conjunt del territori català (Corporació Catalana de Mitjans Audiovisuals, 2023)–.

D'entrada, que la robòtica social s'estigui desenvolupant com a proveïdora d'eines per a contextos professionals fa especialment urgent la reflexió ètica sobre aquest camp tecnocientífic. En efecte, no és el mateix si els robots es conceben, almenys primàriament, per a un context pràctic institucional, o no. Per exemple, no és el mateix un robot com *Harmony* –creat per l'empresa Abyss Creations-Real Doll per a finalitats sexuals[7] i disponible al mercat per a ser adquirit a títol personal per a l'ús privat–, que un robot com l' esmentat *Socrates* –dissenyat expressament com a eina de suport a una activitat terapèutica institucionalitzada. Des d'un punt de vista ètic, és la robòtica social com a facilitadora de recursos per a les institucions que urgeix especialment a una reflexió critico-normativa[8].

No obstant, més enllà de la urgència de la reflexió ètica plantejada, en certa manera, per qualsevol perspectiva d'adopció de noves tecnologies en contextos pràctics professionals, la robòtica social interpel·la èticament en un sentit més fonamental, degut al caràcter inèdit i les dimensions del seu potencial de reconfiguració de la nostra vida. Aquesta branca de la robòtica intel·ligent permet externalitzar part de la nostra agència en tasques que pròpiament només podien fer els humans perquè –i aquí està el nucli de la novetat– requereixen d'una certa interacció personal. Per extensió, la robòtica social obre les portes a introduir mediacions tecnològiques en la dimensió més nuclear de pràctiques de naturalesa relacional, com la cura. La novetat no està en la mateixa mediació tecnològica de tals pràctiques, sinó en una transformació substantiva del caràcter d'aquesta mediació, arran de la capacitat dels robots d'interactuar amb els humans com a quasi-altres. La nova relació d'alteritat humà-tecnologia facilitada per la robòtica social, i caracteritzada per la intencionalitat recíproca de la interacció [9] que la distingeix de la seva versió clàssica (Ihde, 1990), permet estendre la tecnificació a un dels últims bastions d'agència humana que quedava en els àmbits d'activitat relacional.

D'aquí que la particular modificació dels marges d'acció humana per part de la robòtica social posi en qüestió el mateix significat i conceptualització de la pràctica de cura, en el marc d'una controvèrsia sobre si aquesta tecnologia entra en disputa amb el paradigma de les "3Ds" –de l'anglès *dirty, dull* i *dangerous*– sota el qual els robots s'han concebut tradicionalment per tasques que són brutes, avorrides i perilloses (Lin, 2014); o si, per contra, el que està en joc és una certa idea de cura (Liedo i Ausín Díez, 2022). Segons les narratives subjacents al

[7] Gràcies al sistema d'IA, el robot *Harmony* té la capacitat de "conversar" amb les persones i simular emocions. La pretensió de l'empresa és que aquesta capacitat d'interacció social li permeti acomplir finalitats que van més enllà de les pròpiament sexuals, i pugui desenvolupar també un rol més complet com a companya sentimental. Una qüestió èticament controvertida és que, a través d'una aplicació, l'usuari pot ajustar diferents aspectes de la "personalitat" d'aquest robot. Per a una documentació sobre com és *Harmony* en acció, resulta molt il·lustrativa l'obra cinematogràfica *Hi, A.I.* (Willinger, 2019), traduïda al castellà com a *Robots: las historias de amor del futuro*.

[8] Això no implica que el desenvolupament de robots socials per a altres demandes de mercat vinculades a un ús privat en contextos d'activitat no professional no requereixin d'una anàlisi ètica. Tanmateix, la robòtica social destinada a proveir instruments per pràctiques institucionals té prioritat com a objecte de reflexió, tant per l'escala d'implementació tecnològica associada a la mateixa, com per la necessitat d'alinear-ne el desplegament amb els valors i finalitats propis d'aquests contextos pràctics.

[9] Dec aquesta conceptualització a les converses mantingudes amb Peter-Paul Verbeek a la Universitat de Twente durant el tercer trimestre del curs acadèmic 2022.

desplegament de la robòtica social, la cura s'entén com a conjunt de tasques axiològicament diferenciables, algunes de les quals es poden delegar en els robots per ser tasques pesades, repetitives, avorrides i/o mecàniques (Vallès-Peris and Domènech, 2020) –com seria el cas de donar de menjar, ajudar a vestir o fins i tot assistir en exercicis per al manteniment o restauració de la salut física i cognitiva–. És precisament en base a aquesta concepció que des de l'enginyeria (Fundació Víctor Grífols i Lucas, 2019), però també des de les institucions que en promouen el seu desplegament, la robòtica social es defensa com a recurs per a la qualitat de la cura, al permetre descarregar els professionals de salut de tasques menys significatives, en termes de valor humà, i possibilitar que es dediquin a aquelles més definitòries del que seria la "nurturant care" (Duffy, 2007, p.318), i que tenen a veure amb la dimensió més intersubjectiva de les relacions de cura (Generalitat de Catalunya, 2023). En el marc normatiu europeu, per exemple, s'argumenta a favor de destinar els robots a les "tasques automatitzades d'aquells que presten cura [...]" com a manera d' "augmentar la cura humana" (European Parliament, 2017, p.247). Des d'aquesta perspectiva, doncs, la robòtica social mantindria una continuïtat de base amb el model de les 3Ds.

En definitiva, l'entrada dels robots intel·ligents ja no només al nostre món físic sinó també semàntic suposa una disrupció pràctica i estructural –hem de decidir com reorganitzem les nostres activitats de cura a partir de les possibilitats obertes pels avenços en robòtica social–, així com conceptual, al posar en qüestió els significats sobre els quals configurar la nostra vida en comunitat. Aquest és, doncs, el primer dels (dos) sentits en què la robòtica social interpel·la l'ètica, al qual es feia referència a l'inici del capítol. En virtut del seu potencial disruptiu –el significat del qual s'acabarà de desplegar exhaustivament al llarg d'aquest treball–, la robòtica social ens emplaça a una situació de parèntesi: cal pensar quina direcció prenem, i per què. Què, per què, per a què i com: aquests són els grans interrogants que cal afrontar per al desplegament del nou poder que se'ns obre amb la robòtica social, i que demanen d'ètica, en tant que tenen a veure amb la fonamentació dels cursos d'acció.

1.3. Una cartografia de les respostes

En el context del seu creixent desenvolupament, la robòtica social està esdevenint un tema d'atenció ètica central (van Maris *et al.*, 2020). En aquest sentit, podríem parlar, si més no, d'un moviment de resposta a la necessitat d'orientar normativament la força disruptiva de la robòtica social; això és, d'una reacció a la interpel·lació (en el primer dels seus sentits).

D'una banda, a nivell institucional, aquest moviment es veu reflectit en la recent integració de l'ètica en la política europea de desplegament tecnològic en matèria d'IA i robòtica (de Pagter, 2023) –que concerneix, doncs, els sistemes d'IA pertanyents a la robòtica social–. D'entre els diferents actors geopolítics del desenvolupament de la IA –amb EUA, Europa i Xina com a protagonistes principals (Craglia *et al.*, 2018)–, Europa es caracteritza per la seva aposta explícita per una tecnologia al servei de les persones i, conseqüentment, per un compromís amb l'ètica com a principi rector per al seu desenvolupament (Comisión Europea, 2018). D'aquí que s'advoqui per una IA fiable, responsable, centrada en l'humà, beneficiosa: aquestes són les adjectivacions definitòries de la marca tecnològica europea (Grupo de expertos de alto nivel sobre inteligencia artificial, 2019) (European Commission, 2020). Aquest compromís europeu, que sorgeix a partir del 2015 sobretot amb motiu dels riscos d'una creixent autonomia tecnològica dels sistemes d'IA (de Pagter, 2023), s'ha anat cristal·litzant en diferents iniciatives

ètico-normatives per orientar aquesta tecnologia, en les seves diferents aplicacions, de forma afí, coherent i respectuosa amb els valors i drets fonamentals de la Unió Europea. Recentment s'ha consolidat en la seva dimensió legal, amb la primera proposta europea de regulació de la IA, coneguda com *Artificial Intelligence Act* (European Commission, 2021) –en endavant, *AI Act*.

Així doncs, en la seva defensa d'un ideal de tecnologia com a eina per a servir als objectius i valors humans, Europa es fa ressò d'una idea clàssica que, en el cas de la robòtica, queda absolutament fixada per la mateixa etimologia del terme "robot"[10]. Aquesta idea és èticament crucial perquè, com s'exposarà, apunta cap a l'oblidada primacia de la pregunta per la causa final (el 'per a què') en l'activitat de desplegament tecnològic.

Malgrat no figurar-hi encara de forma diferenciada, la robòtica social cau, doncs, dins l'àmbit de consideració de les normatives europees entorn els sistemes d'IA. En primer lloc, com ja s'ha argüit, això es dona per defecte, al ser la robòtica social un camp de la robòtica intel·ligent i compartir certs aspectes èticament rellevants dels sistemes d'IA corporeïtzada. En segon lloc, però, l'atenció es dona també de manera més explícita –tot i que encara tangencialment– amb motiu d'una creixent consideració de les problemàtiques ètiques vinculades a la interactivitat social dels sistemes d'IA[11].

En certa manera, en el document de referència sobre robòtica de la UE –a saber, les *Normes de Dret Civil sobre Robòtica* de 2017 (European Parliament, 2017)– ja s'hi introdueix el que podria considerar-se una primera al·lusió a riscos vinculats amb la capacitat d'interacció social dels robots. Si bé terminològicament no s'usa el concepte de robot social, en aquesta normativa s'assenyala el perill de deshumanitzar la cura que pot suposar la implementació de "robots de cura" o "robots assistencials"[12] en aquesta pràctica, en cas que se substitueixi la interacció social humana per la interacció amb aquests artefactes. En tant que aquests robots es descriuen, a més, com a vinculats, entre d'altres, a funcions d'interacció primàriament social –com fer "companyia" (European Parliament, 2017, p.247)–, s'entén que són robots de tipus social, els que s'hi contemplen. És en aquest sentit que la normativa de 2017 es pot llegir com un primer pas cap a l'emergència de la robòtica social com a objecte d'anàlisi ètic particular.

Posteriorment, en les *Directrius ètiques per a una IA fiable* de 2019, s'introdueix més específicament l'advertència de l'efecte negatiu que la interacció humana amb "sistemes socials d'IA" pot suposar per al benestar físic i mental de les persones, així com per a les competències, relacions i vincles socials humans (High-Level Expert Group on AI, 2019, p.19), i s'insta a la seva avaluació i seguiment. En la *AI Act* aquesta qüestió es torna a formular en relació als riscos que pot suposar la interacció humana amb sistemes d'IA "conversacionals" (European Commission, 2021, p.3). En general, aquests es consideren de risc limitat –un risc contra el qual Europa prescriu transparència, en el sentit de garantir que la persona humana tingui coneixement

[10] Procedent del txec "robota", i usada per primera vegada el 1920 per Karel Čapek en la seva obra R.U.R. Rossumovi Univerzální Roboti, aquesta noció significa "treball forçat" o "servitud".

[11] Aquí s'hi inclouen els sistemes d'IA robòtics i no robòtics.

[12] Resulta significativa la diferència terminològica que hi ha entre la normativa en anglès i la seva versió castellana a l'hora de referir-se a aquests tipus de robots. Mentre que en l'original anglès s'usa el terme "care robots" (European Parliament, 2017, p.247), en la normativa en castellà s'empra el de "robots asistenciales" (Parlamento Europeo, 2017, p.247). Malgrat els motius pels quals es podria donar raó d'aquesta sinonímia, el cas il·lustra l'ambigüitat terminològica generalitzada que existeix entorn els robots socials i que en dificulta l'anàlisi ètic.

d'estar interactuant amb un sistema d'IA–; exceptuant el cas que la interacció estigui destinada expressament a la manipulació, en què llavors els sistemes es cataloguen de risc inacceptable.

D'altra banda, dins la disciplina de Roboètica, la robòtica social –molt particularment en la seva aplicació assistencial– constitueix ja un objecte de reflexió clarament diferenciat d'altres branques de la robòtica intel·ligent (Tzafestas, 2018). Això pressuposa una comprensió de la robòtica social com a camp tecnocientífic que planteja qüestions ètiques específiques, i que reclama ser abordat en la seva particularitat. La proliferació de discussió ètica entorn aquesta tecnologia en el marc de la literatura acadèmica internacional (Vandemeulebroucke, Casterle and Gastmans, 2020) és simptomàtica d'aquest focus –si bé, com es recull en el Capítol 3, la reflexió específica sobre aquesta tecnologia no és sempre fàcil d'identificar, atesa la seva dispersió en un conjunt heterogeni de publicacions i l'ambigüitat terminològica entorn els robots socials (Pareto Boada, Román Maestre and Torras, 2021)–. També n'és un clar indici el creixement que està experimentant la interacció humà-robot (HRI) com a camp de recerca interdisciplinària en vinculació amb la robòtica social, sobretot pel que fa a la seva branca assistencial (Goodrich and Schultz, 2007).

Tanmateix, malgrat la creixent atenció al seu potencial disruptiu i la necessitat d'orientar-lo pertinentment, la interpel·lació a l'ètica que planteja la robòtica social roman, encara, sense resposta. Fonamentalment, això té a veure amb una mancança conceptual i metodològica en relació a l'exercici de reflexió ètica sobre aquesta tecnologia, que dona com a resultat una aproximació ètica predominant deficitària. En part degut a la novetat de la robòtica social, però més substancialment amb motiu d'una vaguetat de fons respecte l'articulació del pensament ètic entorn la tecnologia en general, no es disposa d'un marc de coordenades clar per a l'anàlisi de la robòtica social des d'una perspectiva ètica. En un context en què, a més, la consideració ètica sobre robòtica s'ha liderat sobretot per part de perfils professionals al·lòctons a aquesta branca de la filosofia pràctica (Torras, 2024), això dona lloc a una Roboètica disfuncional, en tant que activitat amb limitacions per a identificar i analitzar críticament les qüestions normativament rellevants per al desplegament de la robòtica social.

Això explica certs trets deficients del panorama de reflexió actual sobre les implicacions ètiques de la robòtica social, com ara la circumscripció del terreny de consideració normativa a l'esfera de la interacció diàdica humà-robot, la dedicació ètica centrada en l'avaluació d'impactes, la descontextualització de la reflexió respecte les pràctiques a què es destinen els robots socials (Pareto Boada, Román Maestre and Torras, 2022)(Pareto Boada, Román Maestre and Torras, 2021), o l'adopció de principis d'altres camps de l'ètica aplicada per a l'avaluació normativa d'aquesta tecnologia (Tzafestas, 2016), generalment de la bioètica –en sintonia amb el que passa amb la tecnologia d'IA en general (Floridi and Cowls, 2019)–.

2. Objectius de recerca

És en el context d'aquesta reacció insuficientment responsiva a la interpel·lació a l'ètica que suposa l'emergència de la robòtica social, que cal entendre el plantejament de la present tesi doctoral. D'entrada, cal recuperar el doble sentit d'aquesta interpel·lació, ja que la investigació recollida en aquest treball es vertebra com a resposta a les dues accepcions del fenomen. D'una

banda pretén proporcionar orientació ètica[13] per al desplegament de la robòtica social, concretament en l'àmbit assistencial. D'altra banda, i com a condició prèvia, es proposa assentar certs fonaments disciplinaris per a l'abordatge ètic d'aquesta tecnologia, és a dir, realitzar una tasca de clarificació i definició de l'ètica en relació a la robòtica social –o sigui, del que designarem com a ètica de la robòtica social[14].

Així doncs, davant la necessitat de desenvolupar una tasca de reflexió ètica rigorosa entorn aquesta nova branca de la robòtica intel·ligent, i tenint en compte el caràcter de les limitacions actuals per a dur-la a terme –que tenen a veure amb l'articulació disciplinària de l'ètica de la robòtica social–, es defineixen dos objectius principals de recerca:

El primer és establir un marc conceptual per a una correcta i exhaustiva aproximació ètica a la robòtica social, especialment en vistes al seu desplegament per a l'àmbit assistencial –això és, la robòtica social assistencial–. Aquest marc s'entén com a teixit conceptual en base al qual identificar i analitzar les qüestions normativament rellevants a l'hora de desenvolupar aquesta tecnologia i fonamentar la presa de decisions.

En vistes a aquest primer propòsit general, la investigació doctoral s'estructura entorn dos objectius secundaris:

- *Objectiu secundari (1).1*: Assentar les bases per a orientar el pensar ètic entorn la robòtica social, és a dir, fonamentar l'activitat de reflexió crítico-normativa sobre aquesta tecnologia. Es pretén clarificar la qüestió *"Què significa (i com) aproximar-se a la robòtica social des d'una perspectiva ètica?"*.

- *Objectiu secundari (1).2:* Delimitar unes coordenades ètiques bàsiques per a orientar el desplegament de la robòtica social assistencial.

El segon dels objectius principals és (re)analitzar qüestions ètiques centrals de la robòtica social assistencial a partir del marc conceptual establert. S'espera contribuir a enriquir la reflexió critico-normativa actual sobre aquesta tecnologia i el tipus de consideracions que han d'acompanyar-ne el seu desplegament.

En aquesta línia, la recerca es desplega en dos objectius secundaris:

[13] Atès el caràcter de l'ètica, aquesta orientació serà sempre mediata, i no s'ha de confondre, per tant, amb el proveïment d'unes directrius compromeses amb determinats cursos d'acció –que indicarien què fer i proporcionarien, en aquest sentit, una orientació immediata–. Aquest punt s'exposarà amb profunditat en el Capítol 2.

[14] A partir d'ara, s'usarà el terme "Roboètica" en el seu sentit disciplinari original, és a dir, per fer referència a l'activitat d'ètica aplicada a la robòtica tal i com ha estat formulada i practicada històricament, i que es constitueix a mans dels enginyers. En canvi, s'anomenarà "ètica de la robòtica" a una activitat articulada des de la disciplina filosòfica, desenvolupada en base als recursos conceptuals i procediments analítics que la defineixen. Tanmateix, això no s'ha d'entendre com una esmena a la necessitat d'interdisciplinarietat reivindicada des de l'inici per la Roboètica: en efecte, cal una col·laboració ininterrompuda entre la filosofia i l'enginyeria. Això és imprescindible per acotar les qüestions normativament rellevants amb coneixement de causa, i respondre-hi pertinentment d'acord amb els marges d'acció de què es disposa en cada cas, (re)vertebrant, així, l'ètica i la tecnologia –que, com s'explicarà al Capítol 2, no són activitats contraposades–.

- *Objectiu secundari (2).1:* Analitzar críticament l'aproximació ètica actual a la robòtica social assistencial.

- *Objectiu secundari (2).2:* Identificar línies d'investigació o temàtiques que convindria desenvolupar per a reorientar l'aproximació actual a la robòtica social assistencial i definir un horitzó disciplinar d'indagació ètica sobre aquesta tecnologia –és a dir, una agenda de recerca de l'ètica en relació a aquest camp de la robòtica intel·ligent–.

A més, en el context d'un retard nacional considerable pel que fa a la incorporació de formació en ètica en els programes universitaris de ciència i tecnologia a Catalunya i Espanya –en comparació amb països europeus com Holanda, així com amb els Estats Units (Mitcham, 2009)– i de les recents manifestacions d'interès i iniciatives per a redreçar-lo (Torras and Ludescher, 2023), s'ha definit un objectiu complementari de recerca. Aquest consisteix en desenvolupar una proposta docent per a una assignatura d'ètica de la tecnologia (6ECTS) per a perfils professionals del camp de l'enginyeria.

3. Continguts i estructura de la tesi

Atès que gran part de la recerca s'ha materialitzat en articles acadèmics, la present tesi es configura com a compendi de publicacions, les quals es complementen substantivament amb altres apartats. Concretament, la tesi s'estructura en dos capítols centrals (Cap. 2, 3), acompanyats del present capítol introductori i un darrer de caràcter conclusiu.

El **Capítol 2**, de caràcter eminentment propedèutic, s'ocupa de definir l'articulació de l'ètica en relació a la robòtica social, a través d'una disquisició disciplinària sobre l'ètica de la tecnologia. D'aquí que aquest capítol concerneixi a *la pregunta per l'ètica de la robòtica social*. En ell es desenvolupa una tasca de fonamentació de l'abordatge de la robòtica social des de l'ètica, que passa per clarificar una qüestió principal: cal dilucidar la significació ètica de la tecnologia, o, altrament dit, el sentit en què la tecnologia –i, en el cas que ens ocupa, la robòtica social– constitueix un objecte per a l'ètica. Això implica reconsiderar la interpel·lació que la tecnologia planteja a l'ètica i que, sobretot en el marc de l'abanderament europeu per una tecnologia al servei de la humanitat, es dona acríticament com entesa. Per què la tecnologia demana d'ètica, exactament?: aquesta és una primera qüestió que requereix conceptualitzar adequadament tant l'ètica com la tecnologia, i que és decisiva per a precisar el tipus d'articulació disciplinar entre ambdues.

El **Capítol 3** consisteix en un recull de quatre articles acadèmics publicats com a part de la investigació doctoral, dels quals se n'adjunta una còpia completa del text. Per a la seva inclusió en aquest treball, i bàsicament en vistes a una uniformitat en l'estil de citació i la configuració de pàgina, s'han realitzat canvis d'edició menors (merament formals).

La referència completa de les publicacions s'indica en una nota a peu destacada en la primera pàgina de cadascun dels articles compilats. Donada la temàtica de la tesi, les revistes seleccionades per a les publicacions pertanyen a diferents àmbits disciplinaris.

A continuació, es descriuen, en línies generals, els diferents articles acadèmics que composen el Capítol 3 de la tesi, d'acord amb el seu ordre cronològic.

El primer article, titulat "Prolegómenos a una ética para la robótica social", és una indagació teòrica que respon a la necessitat de clarificar com orientar el pensament ètic entorn la robòtica social. A partir de l'explicitació i anàlisi de les complexitats pròpies de la deliberació ètica sobre aquesta tecnologia en particular, l'article estableix els elements essencials per a enfocar la reflexió pertinentment.

El segon article, "The ethical issues of social assistive robotics: A critical literature review", consisteix en una anàlisi exhaustiva de l'estat de la reflexió ètica entorn la robòtica social assistencial, que es desenvolupa a partir d'una revisió crítica de la literatura acadèmica sobre els problemes ètics vinculats a aquesta tecnologia. En base als resultats, s'identifiquen i examinen certes tendències de la reflexió que són simptomàtiques d'una aproximació ètica deficient a aquesta branca de la robòtica social.

El tercer article, "Ethics for social robotics: A critical analysis", de caràcter sintètic, ve a complementar l'anàlisi crític sobre l'aproximació ètica actual a la robòtica social per a la cura. Prenent com a representatiu el cas de la reflexió sobre la robòtica social assistencial, l'article en discuteix tres tendències característiques d'una aproximació impròpia d'una ètica de la tecnologia, i defineix línies d'investigació a desenvolupar per redreçar-la.

El quart article, "Social assistive robotics: An ethical and political inquiry through the lens of freedom", corresponent a una de les línies d'investigació formulades en l'article anterior, és una reexaminació filosòfica de les qüestions normativament rellevants per al desplegament de la robòtica social assistencial des del punt de vista de la llibertat humana. L'article posa al centre d'atenció les implicacions per la llibertat vinculades a la dimensió estructural de les relacions humà-robot, la qual roman predominantment desatesa en el marc de discussió actual limitat a la esfera de la HRI.

El compendi d'articles s'acompanya d'un breu **Informe final dels resultats** on s'exposen, de forma concisa, les idees centrals que poden extreure's d'aquesta secció de la tesi, i d'un **Epíleg** de caràcter íntegrament propositiu, on es formulen un parell de línies per a l'agenda de l'ètica de la robòtica social. En concret, es defineixen unes coordenades ètico-polítiques a explorar per al desplegament de la robòtica social assistencial, i s'articula una proposta docent per una assignatura d'ètica de la tecnologia de 6 ECTS dirigida a professionals de l'enginyeria.

Finalment, a la secció de **Conclusions**, es destacaran les contribucions teòriques i pràctiques de la investigació doctoral.

Referències

Andriella, A., Torras, C. and Alenyà, G. (2020) 'Cognitive System Framework for Brain-Training Exercise Based on Human-Robot Interaction', *Cognitive Computation*. doi: 10.1007/s12559-019-09696-2.

Bekey, G. A. (2014) 'Current Trends in Robotics: Technology and Ethics', in Lin, P., Abney, K., and Bekey, G. A. (eds) *Robot Ethics. The Ethical and Social Implications of Robotics*. MIT Press, pp. 17–34.

Breazeal, C., Takanishi, A. and Kobayashi, T. (2008) 'Social Robots that Interact with People', in Siciliano, B. and Khatib, O. (eds) *Springer Handbook of Robotics*. Springer, pp. 1349–1369. doi:

https://doi.org/10.1007/978-3-540-30301-5_59.

Brey, P. (2010) 'Philosophy of Technology after the Empirical Turn', *Techné*, 14(1), pp. 36–48.

Caleb-Solly, P. (2016) 'A brief introduction to … Assistive robotics for independent living', Perspectives in Public Health, 136(2), pp. 70–72.

Caleb-Solly, P. et al. (2014) 'A mixed-method approach to evoke creative and holistic thinking about robots in a home environment', ACM/IEEE International Conference on Human-Robot Interaction, pp. 374–381. doi: 10.1145/2559636.2559681.

Coeckelbergh, M. (2020) *Introduction to Philosophy of Technology*. Oxford University Press.

Coeckelbergh, M. (2022) *Robot Ethics*. The MIT Press.

Comisión Europea (2018) *Comunicación de la Comisión al Parlamento Europeo, al Consejo Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones. Plan coordinado sobre inteligencia artificial*.

Corporació Catalana de Mitjans Audiovisuals (2023) 'ARI, el robot que facilita tasques a la gent gran i arribarà a un miler de llars catalanes', 24 November. Available at: https://www.ccma.cat/324/ari-el-robot-que-facilita-tasques-a-la-gent-gran-i-arribara-a-un-miler-de-llars-catalanes/noticia/3262661/.

Cortina, A. (2003) 'El quehacer público de las éticas aplicadas: ética cívica transnacional', in Cortina, A. and García-Marzá, D. (eds) *Razón pública y éticas aplicadas. Los caminos de la razón práctica en una sociedad pluralista*. Tecnos, pp. 13–44.

Craglia, M. *et al.* (2018) *Artificial Intelligence - A European perspective*. Luxembourg: Joint Research Center. doi: 10.2760/11251.

Didier, C. and Heriard-Dubreuil, B. (2005) 'Engineering Ethics in Europe', in Mitcham, C. (ed.) *Encyclopedia of Science, Technology and Ethics*. Macmillan Reference USA, pp. 632–635.

Dolic, Z., Castro, R. and Moarcas, R. (2019) *Robots in healthcare: a solution or a problem?, Study for the Committee on Environment, Public Health, and Food Safety*, *European Parliament*.

Duffy, M. (2007) 'Doing the dirty work: Gender, race, and reproductive labor in historical perspective', *Gender and Society*, 21(3), pp. 313–336. doi: 10.1177/0891243207300764.

European Commission (2020) *White Paper on Artificial Intelligence - A European approach to excellence and trust*. doi: 10.1017/CBO9781107415324.004.

European Commission (2021) 'Proposal for a Regulation of the European Parliament and of the Council. Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts'.

European Parliament (2017) *Civil Law Rules on Robotics. European Parliament resolution of 16 February 2017 with recomendations to the Commission on Civil Law Rules on Robotics (2015/2013(INL))*, *Official Journal of the European Union*. Available at: http://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html#title10.

Feenberg, A. (2018) 'What Is Philosophy of Technology?', in Beira, E. and Feenberg, A. (eds) *Tecnology, Modernity, and Democracy. Essays by Andrew Feenberg*. Rowman & Littlefield International.

Floridi, L. and Cowls, J. (2019) 'A Unified Framework of Five Principles for AI in Society', *Harvard Data Science Review*, (1), pp. 1–13. doi: 10.1162/99608f92.8cd550d1.

Franssen, M., Lokhorst, G.-J. and van de Poel, I. (2023) 'Philosophy of Technology', *The Stanford Encyclopedia of Philosophy (Spring 2023 Edition)*. Edward N. Zalta & Uri Nodelman (eds.). Available at: https://plato.stanford.edu/archives/spr2023/entries/technology/.

Fundació iSocial (2020) *Misty II, robot per millorar la qualitat de vida de persones grans que viuen soles*.

Fundació Víctor Grífols i Lucas (2019) *'Un robot permet que el cuidador tingui més temps per fer tasques amb valor emocional'*. Available at: https://www.fundaciogrifols.org/ca/-/entrevista-a-carme-torras (Accessed: 20 November 2023).

Generalitat de Catalunya (2023) *El robot que dona de menjar a pacients al programa de Els Matins de TV3*. Available at: https://perevirgili.gencat.cat/ca/detalls/Noticia/El-robot-que-dona-de-menjar-a-pacients-al-programa-de-Els-Matins-de-TV3 (Accessed: 30 November 2023).

Goodrich, M. A. and Schultz, A. C. (2007) 'Human-Robot interaction: A Survey', *Foundations and Trends in Human-Computer Interaction*, 1(3), pp. 203–275. doi: 10.1561/1100000005.

Gordon, J.-S. and Nyholm, S. (2023) 'Ethics of Artificial Intelligence', *The Internet Encyclopedia of Philosophy*. Available at: https://iep.utm.edu/ethics-of-artificial-intelligence/.

Grupo de expertos de alto nivel sobre inteligencia artificial (2019) *Directrices éticas para una IA fiable*. doi: 10.2759/14078.

Heidegger, M. (2009) 'The Question Concerning Technology', in Kaplan, D. M. (ed.) *Readings in the Philosophy of Technology*. 2nd edn. Rowman & Littlefield Publishers, Inc., pp. 9–24.

Heidegger, M. (2021) *La pregunta por la técnica*. 1ª. Herder.

High-Level Expert Group on AI (2019) 'Ethics Guidelines for Trustworthy AI'. European Commission, pp. 1–41.

High-Level Expert Group on Artificial Intelligence (2019) 'A definition of AI: Main capabilities and scientific disciplines. Definition developed for the purpose of the AI HLEG's deliverables'. European Commission, p. 7. Available at: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56341.

Ihde, D. (1990) *Technology and the Lifeworld*. Indiana University Press.

Jonas, H. (2015) *El principio de responsabilidad. Ensayo de una ética para la civilización tecnológica.* Herder. Barcelona.

Kapp, E. (1877) *Grundlinien einer Philosophie der Technik. Zur Entstehungsgeschichte der Kultur aus neuen Gesichtspunkten*. Reprint. 2015 Hamburg: Felix Meiner Verlag.

Kettner, M. (2003) 'Tres Dilemas Estructurales de la Ética Aplicada', in Cortina, A. and García-Marzá, D. (eds) *Razón pública y éticas aplicadas. Los caminos de la razón práctica en una sociedad pluralista*. Tecnos, pp. 145–158.

Liedo, B. and Ausín Díez, T. (2022) 'Alcance y límites de la tecnologización del cuidado: aprendizajes de una pandemia', *Revista Espanola de Salud Pública*, 96.

Lin, P. (2014) 'Introduction to Robot Ethics', in Lin, P., Abney, K., and Bekey, G. A. (eds) *Robot Ethics. The Ethical and Social Implications of Robotics*. The MIT Press, pp. 3–15.

van Maris, A. *et al.* (2020) 'Designing Ethical Social Robots—A Longitudinal Field Study With Older Adults', *Frontiers in Robotics and AI*, 7(January). doi: 10.3389/frobt.2020.00001.

Mejia, C. and Kajikawa, Y. (2017) 'Bibliometric Analysis of Social Robotics Research: Identifying

Research Trends and Knowledgebase', *Applied Sciences (Switzerland)*, 7(12). doi: 10.3390/app7121316.

Mitcham, C. (1994) *Thinking Through Technology. The Path between Engineering and Philosophy*. The University of Chicago Press.

Mitcham, C. (2009) 'A historico-ethical perspective on engineering education: From use and convenience to policy engagement', *Engineering Studies*, 1(1), pp. 35–53. doi: 10.1080/19378620902725166.

Mitcham, C. and Briggle, A. (2009) *The Interaction of Ethics and Technology in Historical Perspective*, *Philosophy of Technology and Engineering Sciences*. Elsevier B.V. doi: 10.1016/B978-0-444-51667-1.50045-8.

Nordmann, A. and Rip, A. (2009) 'Mind the gap revisited', *Nature Nanotechnology*. Nature Publishing Group, 4, pp. 273–274. doi: 10.1038/nnano.2009.26.

Operto, F. and Veruggio, G. (2008) 'Roboethics: Social and Ethical Implications of Robotics', in Siciliano, B. and Khatib, O. (eds) *Springer Handbook of Robotics*. Springer, pp. 1499–1524. doi: 10.1007/978-3-540-30301-5.

Ortega y Gasset, J. (2004) *Meditación de la técnica y otros ensayos sobre ciencia y filosofía*. 8ª. Revista de Occidente en Alianza Editorial.

de Pagter, J. (2023) 'From EU Robotics and AI Governance to HRI Research: Implementing the Ethics Narrative', *International Journal of Social Robotics*. doi: 10.1007/s12369-023-00982-6.

Pareto Boada, J., Román Maestre, B. and Torras, C. (2021) 'The ethical issues of social assistive robotics: A critical literature review', *Technology in Society*, 67. doi: 10.1016/j.techsoc.2021.101726.

Pareto Boada, J., Román Maestre, B. and Torras, C. (2022) 'Ethics for social robotics: A critical analysis', in *TRAITS Workshop Proceedings (arXiv:2206.08270) held in conjunction with Companion of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*. Springer Berlin Heidelberg, pp. 1284–1286.

Parlamento Europeo (2017) 'Normas de Derecho Civil sobre robótica. Resolución del Parlamento Europeo, de 16 de febrero de 2017, con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica (2015/2013(INL))'. Diario Oficial de la Unión Europea.

van der Plas, A., Smits, M. and Wehrmann, C. (2010) 'Beyond speculative robot ethics: A vision assessment study on the future of the robotic caretaker', *Accountability in Research*, 17(6), pp. 299–315. doi: 10.1080/08989621.2010.524078.

Plató (1988) *Diàlegs, vol. IX*. Barcelona: Fundació Bernat Metge.

Rabbitt, S. M., Kazdin, A. E. and Scassellati, B. (2015) 'Integrating socially assistive robotics into mental healthcare interventions: Applications and recommendations for expanded use', *Clinical Psychology Review*. Elsevier B.V., 35, pp. 35–46. doi: 10.1016/j.cpr.2014.07.001.

Sætra, H. S. and Danaher, J. (2022) 'To Each Technology Its Own Ethics : The Problem of Ethical Proliferation', *Philosophy & Technology*. Springer Netherlands, 35(93), pp. 1–26. doi: 10.1007/s13347-022-00591-7.

Sarrica, M., Brondi, S. and Fortunati, L. (2020) 'How many facets does a "social robot" have? A review of scientific and popular definitions online', *Information Technology and People*, 33(1), pp. 1–21. doi: 10.1108/ITP-04-2018-0203.

Torras, C. (2019) 'Assistive Robotics : Research Challenges and Ethics Education Initiatives', *Dilemata, Revista Internacional de Éticas Aplicadas*, (30), pp. 63–77.

Torras, C. (2024) 'Ethics of Social Robotics: Individual and Societal Concerns and Opportunities', *Annual Review of Control, Robotics, and Autonomous Systems*, 7(1), pp. 1–18. doi: 10.1146/annurev-control-062023-082238.

Torras, C. and Ludescher, L. G. (2023) 'Writing Science Fiction as an Inspiration for AI Research and Ethics Dissemination', Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 13500 LNAI, pp. 322–344. doi: 10.1007/978-3-031-24349-3_17.

Tzafestas, S. G. (2016) 'Socialized Roboethics', in *Roboethics. A Navigating Overview*. Springer.

Tzafestas, S. G. (2018) 'Roboethics: Fundamental concepts and future prospects', *Information (Switzerland)*, 9(6). doi: 10.3390/INFO9060148.

United Nations Department of Economic and Social Affairs (2022) *World Population Prospects 2022. Summary of Results*. Available at: www.un.org/development/ desa/pd/.

Vallès-Peris, N. and Domènech, M. (2020) 'Roboticists' Imaginaries of Robots for Care: The Radical Imaginary as a Tool for an Ethical Discussion', *Engineering Studies*, 12(3), pp. 157–176. doi: 10.1080/19378629.2020.1821695.

Vandemeulebroucke, T., Casterle, B. D. and Gastmans, C. (2020) 'Ethics of socially assistive robots in aged-care settings: A socio-historical contextualisation', *Journal of Medical Ethics*, 46(2), pp. 128–136. doi: 10.1136/medethics-2019-105615.

Verbeek, P.-P. (2010) 'Accompanying Technology: Philosophy of Technology after the Ethical Turn', *Techne: Research in Philosophy and Technology*, 14(1), pp. 49–54.

Verbeek, P.-P. (2011) *Moralizing Technology: Understanding and Designing the Morality of Things*. The University of Chicago Press.

Veruggio, G. (2006) *EURON Roboethics Roadmap*.

Veruggio, G. and Operto, F. (2006) 'Roboethics: A bottom-up interdisciplinary discourse in the field of applied ethics in robotics', *International Review of Information Ethics*, 6, pp. 2–8. doi: 10.4324/9781003074991-9.

Veruggio, G., Solis, J. and Van Der Loos, M. (2011) 'Roboethics: Ethics applied to robotics', *IEEE Robotics and Automation Magazine*, 18(1), pp. 21–22. doi: 10.1109/MRA.2010.940149.

Willinger, I. (2019) *Hi, A.I.* Available at: https://www.filmin.cat/pelicula/hi-a-i.

CAPÍTOL 2. LA PREGUNTA PER L'ÈTICA DE LA ROBÒTICA SOCIAL

1. Consideracions propedèutiques

Assentar els fonaments per a una correcta articulació de l'ètica en relació a la robòtica social passa, en primer lloc, per atendre dues preguntes intrínsecament relacionades: en quin sentit la tecnologia demana d'ètica?; i de quina ètica es tracta? D'aquí que, com a propedèutica a l'ètica de la robòtica social, aquesta secció comenci per oferir una conceptualització de l'ètica i de la tecnologia que permeti dilucidar el sentit de la seva relació i, en conseqüència, la manera com aquesta ha de conjugar-se disciplinàriament.

1.1. Sobre l'ètica

Sense pretendre oferir una introducció exhaustiva a l'ètica com a disciplina filosòfica, a continuació se n'assenyalen alguns trets que són essencials per a comprendre-la en la seva interacció amb la tecnologia.

D'entrada, des d'una perspectiva filosòfica, entendre l'ètica pròpiament implica distingir-la de la moral. Tot i que 'ètica' (del grec *ethos*) i 'moral' (del llatí *mos-mores*) són etimològicament equivalents i s'usen popularment de forma indistinta, com a termes filosòfics es distingeixen contemporàniament pel que fa a la seva connotació (Ricoeur, 2008) (Cortina and Martínez, 2001). En un moment històric en què l'ètica no només està en voga, sinó que se la sol·licita proactivament per a donar suport en la presa de decisions en diferents contextos pràctics, demarcar amb exactitud aquesta distinció resulta important, sobretot a fi d'evitar malentesos interdisciplinaris. I és que, si bé és cert que tant l'ètica com la moral tenen a veure amb l'exercici d'orientar l'acció, ho fan de forma substantivament diferent.

Atenent a la distinció filosòfica, la moral fa referència a l'adhesió a un conjunt de valors i regles a partir dels quals estimem les accions com a bones o dolentes. En aquest sentit, la noció de 'moral' mantindria l'accepció etimològica originària tant del llatí com del grec, a saber: 'hàbit' o 'costum'. En canvi, l'ètica, com a disciplina normativa, consisteix en l'activitat de reflexió crítico-racional sobre la moral, això és, en l'exercici de la seva fonamentació (Román Maestre, 2016). D'aquí que, a grans trets, a l'ètica se la conceptualitzi disciplinàriament com a estudi sobre la moral, i se la designi comunament com Filosofia moral (López Aranguren, 1991)(Cortina, 2003). En aquest sentit, l'ètica es defineix també, més concretament, com una reflexió de segon grau sobre la moral –això és, una "meta moral" (Ricoeur, 2008, p.48)–, en tant que s'ocupa de la seva legitimitat.

És a aquesta distinció de grau o nivell que fa referència la coneguda contraposició filosòfica entre la moral com a *viscuda* i la ètica com a *pensada* (López Aranguren, 2005), altrament sintetitzada a través de la següent formulació: mentre que la moral respon a la pregunta "què he de fer?" –i la resposta és una acció o omissió –, l' ètica contesta a la pregunta "per què haig de?" –essent, la resposta, un argument– (Cortina, 2007, p.62).

Així doncs, si bé l'ètica, com la moral, orienta l'esfera de les accions humanes –motiu pel qual se la reclama, precisament, en els diferents àmbits d'activitat humana–, ho fa d'una manera distintiva. Mentre que la moral orienta de forma immediata, l' ètica ho fa de forma mediata, a través d'una tasca de fonamentació dels principis de l'acció moral (Cortina, 2003). Amb això

queda clar que ètica i moral s'ocupen, en el fons, de dues dimensions de la nostra vida pràctica que és imprescindible que distingim: les accions, d'una banda, i les raons, de l'altra.

Lluny de dictaminar cursos d'acció humana, l'ètica, com a branca de la filosofia, s'ocupa de les raons, de la fonamentació, de l'àmbit del *per què* de les accions –i, més generalment, de la conducta humana–. És en aquesta justa mesura, doncs, que cal entendre el caràcter normatiu de l'ètica (López Aranguren, 1991). Aquest és un primer tret essencial de l'ètica que cal tenir present en el marc del seu desplegament contemporani com a ètica aplicada, és a dir, com a filosofia moral en relació a activitats humanes específiques –entre elles, l'enginyeria robòtica.

Aquí, resulta útil fer un incís taxonòmic sobre l'ètica per a clarificar com caldrà entendre exactament la modalitat pròpia de l'ètica de la tecnologia. Tradicionalment, s'acostuma a distingir entre diferents formes d'ètica, que vindrien a correspondre a diferents nivells de l'activitat critico-racional, i que s'han de comprendre també en la seva vinculació amb períodes concrets de la història de la filosofia. A saber: l'ètica normativa[15], l'ètica aplicada i la metaètica. D'una banda, quan l'activitat de fonamentació de la moral es manté a un nivell primàriament teòric, en el sentit que s'ocupa de definir els criteris en base als quals valorar i fonamentar la correcció i incorrecció de les accions, ens trobem davant el que seria l'ètica normativa genèricament entesa. És en relació a aquest procés de fonamentació que es desenvolupen els tres grans models ètics clàssics des dels quals articular la moralitat –el teleològic, el deontològic i el de les virtuts–, que, sobretot en la seva complementarietat, segueixen sent un marc de referència per a la Filosofia moral contemporània (Camps, 2017). D'altra banda, quan la tasca de fonamentació de l'acció moral es desenvolupa en relació a un camp d'activitat o pràctica humana específica – circumscrivint-se la reflexió, doncs, a les problemàtiques morals concretes que hi apareixen–, parlem d'ètica aplicada. En tercer lloc, quan la reflexió se centra, més fonamentalment, en l'estudi de la pròpia significació del llenguatge i termes morals, ens trobem davant del que es coneix com a metaètica.

Com s'ha avançat en el capítol anterior, la reflexió ètica específica sobre la tecnologia –i, més concretament, sobre les tecnologies– es demarca disciplinàriament com a tasca pròpia de l'ètica aplicada (Franssen, Lokhorst and van de Poel, 2023) (Gordon and Nyholm, 2023) (Sætra and Danaher, 2022). Quina sigui la idoneïtat de tal conceptualització és una qüestió que es tractarà més endavant. Per ara, cal assenyalar que, en la seva comesa eminentment pràctica, els diferents nivells d'activitat critico-racional de l'ètica no es poden desvincular, en última instància, l'un de l'altre (Bayertz, 2003). Aquest és un segon tret clau per a l'articulació de l'ètica en relació a àmbits pràctics concrets, que esdevé urgent recordar per al cas de l'ètica de la tecnologia. En l'anàlisi de qüestions morals vinculades a la tecnologia s'hi conjugaran moments reflexius que, des del punt de vista d'una categorització taxonòmica, correspondrien a altres

[15] En el marc de la taxonomia clàssica, l'adjectivació de "normativa" per a la forma més general d'ètica no ha d'interpretar-se en un sentit excloent, doncs, d'acord amb la definició oferta anteriorment, també en la seva forma "aplicada" l'ètica té un caràcter normatiu. Per aquest motiu, si bé en el context de l'ètica de la tecnologia s'ha adoptat la nomenclatura pròpia de la classificació tradicional (Sætra and Danaher, 2022) (Tzafestas, 2016), alternatives terminològiques com ara "ètica general" (Kettner, 2003, p.147), "ètica com a tal" (MacIntyre, 1984, p.499), "ètica tradicional" (Kettner, 2003, p.148), o "ètica filosòfica" són més pertinents per a designar aquesta primera forma d'ètica primàriament teòrica.

formes d'ètica; moments com la revisió crítica del significat de certs valors i principis de la moralitat que es poden veure desdibuixats conceptualment davant les noves tecnologies.

A més, la caracterització corrent de l'ètica de la tecnologia porta a la qüestió sobre el procedir de l'ètica aplicada, que resulta important clarificar en el context contemporani d'una manca de consens pel que fa a l'estatut i mètode d'aquesta disciplina –una mancança que queda manifestada en la literatura acadèmica concernent a l'ètica de la tecnologia–. En què consisteix, exactament, això de l' 'aplicació' de l'ètica? D'entre els diferents models existents, destaca, per la seva consistència no només teòrica sinó també pràctica, el model d'ètica aplicada com a "hermenèutica crítica" desenvolupat per Cortina (1996).

Compromesa amb la necessitat d'orientar l'acció en la vida real, l'ètica atén a diversos àmbits d'activitat humana en la seva especificitat, això és, a les situacions i problemes morals que apareixen en relació a aquests. Per això l'ètica aplicada es defineix sovint com una "ètica de les professions" (Camps, 2017, p.393), si bé la idea d' "activitats socials" entesa en termes macintyreans de 'pràctica' (Cortina, 1996, p.130) resulta més precisa a l'hora de fer referència al camp de l'ètica aplicada pròpiament dita.

Tal i com la defineix MacIntyre, una pràctica és una "actividad humana cooperativa, establecida socialmente, mediante la cual se realizan los bienes inherentes a la misma mientras se intenta lograr los modelos de excelencia que le son apropiados a esa forma de actividad y la definen parcialmente […]"(MacIntyre, 2019, p.233). És a dir, es tracta d'una activitat social caracteritzada per uns determinats "béns interns" (MacIntyre, 2019, p. 234), o sigui, unes finalitats específiques de l'activitat que en defineixen la seva raó de ser –el seu 'per a què' o *telos*–. Per això, les pràctiques són activitats que troben la seva legitimitat en la persecució d'aquests béns interns, els quals, per ser assolits, requereixen de certs hàbits o "virtuts" per part dels que participen en l'activitat (MacIntyre, 2019, p.237). En aquest sentit, suposa corrompre l'activitat el fet de practicar-la perseguint-t'hi finalitats que no en són les específiques (Ruiz Trujillo, 2020), com ara els seus "béns externs" (MacIntyre, 2019, p.234) .

La raó per la qual és més apropiat parlar d' activitats socials que no pas de professions al referir-nos al terreny de l'ètica aplicada entronca amb un requisit fonamental per a una 'aplicació' ben entesa, i que és el que precisament recull el model de l'hermenèutica crítica, distingint-se dels models tradicionals d'ètica aplicada[16]. A saber: la necessitat d'atendre, en el procés de fonamentació de les decisions i accions, a les finalitats (béns interns) i valors que són definitoris de l'activitat humana en qüestió –i que no esgoten els models d'excel·lència associats a la configuració professional de l'activitat[17]–. Altrament dit: l'ètica ha de contextualitzar la reflexió en el marc teleològic i axiològic particular de l'activitat en relació a la qual s'emprèn la tasca de fonamentació moral.

[16] S'entenen com a mètodes tradicionals d'ètica aplicada l'anomenada Casuística 1 i Casuística 2 (Arras, 1990), els quals representen, desencertadament, el model d' 'aplicació' de referència de l'ètica de les tecnologies contemporània.

[17] En l'articulació professional de l'activitat s'hi conjuguen altres finalitats (béns externs) i valors vinculats, per exemple, als mecanismes socials necessaris per a dur a terme l'activitat, com ara el mercat i la competència, que fan que l'eficiència, la recerca del benefici o l'obtenció de finançament, entre d'altres, siguin necessàries per assolir els béns interns de l'activitat.

En els models tradicionals d' 'aplicació' de l'ètica, la fonamentació de les accions procedeix o bé de forma deductiva, a partir de principis morals universals (Casuística 1), o bé de forma inductiva, a partir de màximes d'actuació formulades en base a l'anàlisi de casos concrets (Casuística 2). A diferència d'aquests, el model de l'hermenèutica crítica procedeix a partir d'un anàlisi de les finalitats i valors específics de l'activitat en qüestió que, sempre en consonància amb els principis de l'ètica cívica (justícia i dignitat) i dialògica (reconeixement dels afectats com a interlocutors vàlids) (Cortina, 2003), constituiran el marc a la llum del qual orientar les accions concretes en el si de la pràctica. D'aquí la terminologia "hermenèutica crítica de les activitats humanes" per a definir el model metodològic de l'ètica aplicada (Cortina, 1996).

D'acord amb aquesta primera (i modesta) aproximació a la disciplina, si l'ètica s'ocupa de la fonamentació critico-racional de la moralitat –i de la de cursos d'acció més particularment–, aviat es planteja un interrogant clau: què hi té a veure, la tecnologia, amb la moralitat? Altrament formulat: com entra la tecnologia en joc amb aquest regne que, d'acord amb la metafísica moderna (Latour, 1993), es pressuposa com a territori d'una agència exclusivament humana (Verbeek, 2011)? De la resposta a aquesta pregunta, que depèn de la caracterització de la tecnologia de què es parteixi, se'n derivarà el tipus d'ètica que li correspon.

1.2. Sobre la tecnologia

1.2.1. Una noció polièdrica

Conceptualitzar la tecnologia no és una tasca fàcil. D'una banda, això es deu a la varietat de manifestacions sota les quals es pot donar aquest fenomen: "tecnologia" pot dir-se com a objecte (artefacte), activitat, sistema o infraestructura, coneixement o, fins i tot, volició (Mitcham, 1994). D'altra banda, també hi afegeix complexitat la pròpia accepció històrica del terme "tecnologia", des de la qual se'l pot interpretar com a concernent estrictament al tipus d'activitat i artefactes de la tècnica del s. XX.

A l'efecte de dilucidar el sentit en què la tecnologia concerneix a l'ètica, convé sobretot atendre a la tecnologia en la seva qualitat d'artefacte i, derivadament, d'activitat. Aquestes són, de fet, les dues conceptualitzacions principals en base a les quals s'ha anat definint i modulant històricament l'agenda de reflexió filosòfica sobre la tecnologia (González García and Fernández-Jimeno, 2022)– i, encara més incisivament, la de l'ètica–.

En la seva condició d'objecte, la tecnologia s'entén generalment com a eina, o instrument; això és, com a un mitjà per a uns fins. Aquesta és l'anomenada definició instrumental de la tecnologia (Heidegger, 2009), que va predominar fins al s. XX de la mà del pressupòsit modern de neutralitat axiològica de la tecnologia (Feenberg, 2018), i que ressona encara en la contemporaneïtat (Pitt, 2014), malgrat les sòlides crítiques que ha rebut de part de la Filosofia de la tecnologia, així com del camp dels Estudis de Ciència i Tecnologia (STS). En sintonia amb la dicotomia subjecte-objecte pròpia de la Modernitat, que roman considerablement intacta en el nostre pensament moral (Verbeek, 2011), l'associació sembla evident: en tant que és una eina (objecte), la tecnologia no és ni bona ni dolenta: tot depèn dels usos que se li doni. En efecte, partint de la dicotomia ontològica moderna, la moralitat només pot tenir a veure amb els subjectes racionals,

que són els que tenen llibertat i intenció, mentre que els objectes són, en aquest sentit, passius i muts (Verbeek, 2011)[18].

El "*Guns don't kill people, people kill people"* de l'Associació Nacional del Rifle Americà és un dels predicaments referencials de la concepció instrumental (moderna)[19], d'acord amb la qual els artefactes tecnològics entren en relació amb la moralitat en virtut de les conseqüències del seu ús i de les finalitats que se'n volen assolir.

El tipus d'ètica corresponent a aquesta concepció de la tecnologia es caracteritza per delimitar el seu objecte de reflexió normativa als usos (moralment acceptables o no) dels artefactes –i, com a molt, a la deliberació de les finalitats en relació a les quals la tecnologia es concep primàriament com a mitjà–. En definitiva, és només amb motiu dels usos i els seus efectes (en molts casos, imprevisibles) que podem referir-nos a les tecnologies com a essent susceptibles d'avaluació moral. Al seu torn, això restringeix de forma dràstica el terreny de consideració normativa que concerneix a la tecnologia en la seva condició d'activitat, que s'entén correlativament com la mera producció d'aquests instruments al servei de fins humans.

1.2.2. Assalts filosòfics a la neutralitat axiològica de la tecnologia

La presumpció de neutralitat axiològica que ha acompanyat històricament la definició instrumental de la tecnologia ha estat impugnada insistentment per la Filosofia de la tecnologia, tant clàssica com contemporània (Brey, 2010), sobretot per part de tres corrents: la fenomenologia clàssica, la postfenomenologia i la teoria crítica de la tecnologia.

Un dels primers moviments decisius en aquesta direcció es troba en la reconceptualització no-instrumental de la tecnologia que formula Heidegger ja al 1953[20]: la tecnologia no és només un mitjà per a uns fins, sinó també, i més fonamentalment, "un mode de revelar" (Heidegger, 2009, p.13). Què significa això? En termes senzills, la idea vindria a ser que la tecnologia és el marc en què se'ns dona la realitat –d'aquí que la "reveli"– i, en aquest sentit, condiciona la nostra manera de relacionar-nos-hi. La tecnologia configura la manera en què el món ens apareix (es fa present per a nosaltres) i, al seu torn, la manera com l'entenem i ens hi relacionem. La tècnica de la navegació, per exemple, ens descobreix el mar com a navegable, així com les tècniques agrícoles ens mostren la terra com a cultivable (Esquirol, 2011). Quin sigui el particular mode de revelar de la tecnologia (en la seva accepció de tècnica moderna) és una qüestió subseqüent i extensament analitzada per Heidegger que escapa l'objecte d'aquesta secció. Per a la tasca de conceptualització que ens ocupa, n'hi ha prou amb la contribució que suposa la definició no-

[18] Des de la filosofia de la tecnologia contemporània es critica aquesta separació radical entre subjectes i objectes, pròpia de la metafísica humanista i de l'ètica que predomina actualment. Es critica no perquè es consideri que no hi ha diferència entre els humans i els artefactes, en termes d'agència i responsabilitat moral (Verbeek, 2009), sinó perquè, pels motius que s'exposaran més endavant, cal matisar-la per a integrar correctament el rol moral dels artefactes, el seu paper en l'agència humana i, per tant, la seva rellevància ètica en la reflexió normativa. Més que ser producte del subjecte moral autònom que planteja la Il·lustració, l'agència del subjecte moral està intrínsecament lligada a l'entorn material i els artefactes, que hi juguen un rol decisiu (Verbeek, 2011).

[19] En endavant, s'usarà aquesta expressió per fer referència a la comprensió de la tecnologia com a instrument moralment neutre, per a distingir-la del que serà una definició instrumental apropiada, això és, desvinculada de la tesi de neutralitat axiològica.

[20] Aquesta definició la va plantejar originalment en la conferència *Die Frage nach der Technik,* impartida a l'Acadèmia Bàvara de Belles Arts aquell any.

instrumental de la tecnologia. D'acord amb aquesta, pròpia d'una teoria substantivista de la tecnologia (Feenberg, 2018, p.59), lluny de ser un mer mitjà neutral, la tecnologia és una forma de pensar i veure el món, i implica, doncs, una càrrega axiològica determinada.

En el marc del seu primer gir empíric (1980-1990) (Brey, 2010), la filosofia de la tecnologia contemporània, en particular l'anomenada postfenomenologia (Ihde, 2015) (Rosenberger and Verbeek, 2015), refina la crítica a la neutralitat de la tecnologia. Concretament, ho fa assenyalant el fenomen de la mediació tecnològica; això és, el rol mediador de les tecnologies en la relació dels humans amb el món, tant en termes hermenèutics com existencials (Verbeek, 2005). En el seu ús, els artefactes tecnològics ens donen accés a la realitat en certes formes i no d'altres, i ens conviden a determinades accions i no d'altres, tal i com s'evidencia a partir de l'anàlisi postfenomenològic clàssic de les relacions humà-tecnologia (Ihde, 1990), i de l'estudi del rol dels artefactes en l'acció humana per part dels Estudis de Ciència i Tecnologia (STS) (Latour, 1994), respectivament.

Atesa aquesta doble dimensió de la mediació tecnològica –la mediació de la percepció (o experiència), que es dóna com a reducció o amplificació de certs aspectes de la realitat, i la mediació de l'acció (o existència), que es dóna com a invitació o inhibició de comportaments específics (Verbeek, 2005)(Verbeek, 2011)–, les tecnologies tenen un paper constitutiu en la manera com els humans percebem i interpretem la realitat, així com en les nostres decisions i accions. En definitiva, articulen la nostra forma d'estar en el món. Això explica que les tecnologies es defineixin, postfenomenològicament, com el "medi de l'existència humana" [21] (Rosenberger and Verbeek, 2015) i que des de la teoria crítica de la tecnologia es caracteritzin com a marcs per a modes de vida (Feenberg, 2018). Lluny de ser instruments neutrals que senzillament compleixen les seves funcions –o, altrament dit, mers intermediaris entre els humans i la realitat de què aquests participen–, les tecnologies contribueixen activament a la configuració de les experiències i pràctiques humanes. Just per això tenen una significació moral que va més enllà de la que se li atribueix tradicionalment en virtut de la seva condició instrumental o funcional (amb raó dels efectes que se'n poden derivar o les finalitats a què serveix): participen en la configuració de la nostra condició i circumstància moral, i *medien*, en aquest sentit, la nostra presa de decisions i accions. D'aquí que es parli de mediació tecnològica de la moralitat (Kudina and Verbeek, 2019) (Kudina, 2019).

L'exemple de referència per a il·lustrar aquesta participació activa de les tecnologies en la moralitat és el de l'ecografia obstètrica, exhaustivament analitzada des de la teoria de la mediació (Verbeek, 2008). Contra el que podria semblar d'entrada, aquesta tecnologia d'ultrasò no és simplement una eina que ens permet veure el fetus a l'úter, és a dir, un mitjà que *senzillament* ens hi dóna accés visual i ens permet obtenir-ne informació mèdicament rellevant. Lluny de ser una mera "finestra a l'úter" (Verbeek, 2011, p.24), l'ecografia obstètrica *media* la

[21] L'etimologia de "mediació" (del llatí 'mediare') explica què significa que les tecnologies 'mediïn' les relacions humà-món: les tecnologies se situen enmig de la relació dels humans amb el món, tant pel que fa a la dimensió hermenèutica com pràctica d'aquesta relació. Les tecnologies se situen entre els humans i la realitat amb què aquests es relacionen, però de manera activa, donant forma a la relació: fan present la realitat per als humans d'una forma específica i defineixen la manera particular com els humans estan presents en el món. D'aquí que, des de la postfenomenologia, les relacions humà-món s'hagin redefinit com a "relacions humà-tecnologia-món" –o, més sintèticament, "relacions humà-tecnologia" (Ihde, 1990, p.41).

relació entre el no-nascut i els pares: al fer aparèixer i configurar el fetus d'una manera ontològicament distintiva, com a entitat individual i pacient mèdic (mediació de la percepció), emplaça els pares a determinades situacions de tria en relació a aquest (mediació de l'acció). No es tracta només de què l'ecografia obstètrica situï els pares en la tessitura d'haver de decidir què fer en cas de què l'ecografia reveli problemes de salut greus en el fetus –si avortar o no –, sinó que aquesta tecnologia influeix en les seves decisions pel mer fet mateix d'haver de decidir si sotmetre's a la prova o no. Haver de decidir si fer la prova o no per conèixer potencials riscos de salut en el fetus és ja una decisió a què els emplaça la tecnologia. Fixem-nos que l'ecografia obstètrica resignifica la idea de responsabilitat dels pares pel seu fill no nascut i del que implica fer-se'n càrrec també a nivell social, modificant els marcs culturals i morals d'interpretació. Des del moment en què les afectacions de salut del no-nascut esdevinen una qüestió de tria i no d'atzar, no sotmetre's a l'ecografia obstètrica es converteix en una irresponsabilitat, en tant que assumpció deliberada d'un risc evitable.

Tant la conceptualització no-instrumental de la fenomenologia clàssica com la que ofereix la teoria de la mediació de la postfenomenologia comprenen la tecnologia com a *font* d'una manera particular d'entendre i estar en el món. Això les singularitza respecte d'opugnacions contemporànies a la neutralitat tecnològica que s'argumenten en base a una altra dimensió clau de la tecnologia que, malgrat estar vinculada al seu estatut instrumental, roman desatesa en la mateixa concepció instrumental (moderna), a saber: la dimensió social de la tecnologia, tal i com ha estat particularment evidenciada per la teoria crítica de la tecnologia (Feenberg, 2010).

En tant que producte d'un context social i històric concret i, per tant, de formes de vida i interessos específics, les tecnologies tenen, per defecte, una certa càrrega axiològica. Lluny de ser neutral, la racionalitat funcional de les tecnologies respon i es vincula a una forma específica d'entendre i relacionar-nos amb el món i amb els altres[22]. En efecte, és en virtut d'un particular compromís amb la realitat –això és, a una circumstància existencial particular, definida per unes condicions epistèmiques, antropològiques, morals i polítiques específiques–, que les tecnologies s'erigeixen com a instruments. D'entrada, doncs, en la seva qualitat d'artefactes, les tecnologies incorporen valors propis del substrat hermenèutico-pràctic concret del que resulten i formen part –altrament dit, d'una determinada cosmovisió (Hui, 2020), en el sentit més ample del terme–. Aquesta configuració axiològica de les tecnologies s'ha d'entendre com a incorporant diferents capes: és resultat de la cosmovisió sociohistòrica general des de la qual es desplega la tecnologia, i també de la dels actors que hi participen directament, de la seva (inter)subjectivitat situada. La dimensió social de les tecnologies explica la creixent atenció contemporània a la codificació no deliberada de valors socioculturals moralment injustificables en les tecnologies, en part procedents dels professionals que les desenvolupen.

1.2.3. La dimensió política de la tecnologia

La no neutralitat moral de la tecnologia, i molt particularment el caràcter mediador dels artefactes tecnològics, té una rellevància decisiva en la dimensió política de la tecnologia; això és, el seu rol en la configuració sociopolítica de la vida humana. Aquesta és una altra dimensió

[22] És atenent a la no neutralitat de la racionalitat tecnològica il·luminada per la teoria crítica que, més recentment, en el marc de la filosofia de la tecnologia contemporània es parla de la "tecnodiversitat" (Hui, 2020) com a ideal de configuració sociotècnica del món, en tant que englobaria formes de racionalitat tecnològica que responguin a diferents cosmovisions.

constitutiva de la significació ètica de la tecnologia, que convé clarificar. De fet, fer-ho resulta urgent, atès el context actual d'una ètica de la tecnologia simptomàticament caracteritzada per l'oblit d'aquesta dimensió política (Coeckelbergh, 2022; 2018); un oblit que, com es mostrarà en el Capítol 3, queda ben reflectit en la reflexió normativa a l'entorn de la robòtica social assistencial (Pareto Boada, Román Maestre and Torras, 2021).

El fenomen de mediació tecnològica implica que la dimensió política de la tecnologia no es redueix a la funcionalitat instrumental dels artefactes, com així es pressuposa en la concepció instrumental (moderna). És evident que, per la seva dimensió instrumental, la tecnologia pot servir com a mitjà per a fins polítics concrets. Per tant, pot ser usada deliberadament per a l'exercici del poder i l'organització social, i això independentment de si ha estat concebuda per a aquests o no, i sigui a través de l'ús o de les propietats materials, de disseny, de l'artefacte, com és el cas de l'arquitectura hostil (Rosenberger, 2017). Els ponts sobre les autopistes a Long Island de l'arquitecte R. Moses són un exemple paradigmàtic d'aquesta última forma d'instrumentalitat política dels artefactes, en què les característiques arquitectòniques dels ponts –que tenien una alçada massa baixa per al trànsit del transport públic– impedien que les classes socials baixes i les minories racials, que es desplaçaven amb autobús, poguessin accedir a la Jones Beach (Winner, 2009). Pel seu disseny, els ponts imposaven un ordre social de desigualtats d'oportunitat substantiva.

Tanmateix, la participació de la tecnologia en l'establiment d'un ordre específic de relacions socials no es dona només en virtut de la instrumentalitat política dels artefactes, sinó també, i més fonamentalment, amb motiu del seu caràcter mediador en les relacions humà-món. El rol de les tecnologies en la configuració sociopolítica de la vida humana té a veure, en últim terme, amb el doble nivell *micro* i *macro* de la mediació tecnològica (Verbeek, 2020). Com assenyalen els estudis postfenomenològics més recents, les tecnologies no només co-determinen les percepcions i experiències dels usuaris, i les seves decisions i accions, sinó que també contribueixen a configurar els marcs culturals i morals d'interpretació, i, amb ells, les pràctiques i la organització social (Kudina, 2019)(Kudina and Verbeek, 2019). Amb la configuració de les experiències i interpretacions de la realitat, i, al seu torn, dels respectius marges d'acció i les situacions de tria en aquesta, les tecnologies organitzen relacions de poder específiques, a nivell intersubjectiu *micro* i a nivell de societat, *macro*. És a dir, les tecnologies contribueixen a vertebrar la disposició relacional dels individus en el si de la comunitat. Un cas il·lustratiu de la dimensió política de la tecnologia en aquest sentit és el rol constitutiu de les píndoles anticonceptives en l'estatus social, al desconnectar el sexe de la reproducció i modificar així els marcs interpretatius i normatius sobre la sexualitat (Kudina, 2019).

Al matisar una dimensió política de la tecnologia que va més enllà de la funcionalitat instrumental dels artefactes, la mediació tecnològica explica alhora que les tecnologies tinguin implicacions polítiques no intencionades; és a dir, un rol en la constitució de l'ordre sociopolític més enllà del que es pretén deliberadament, com passa amb l'arquitectura estàndard, d'accessibilitat no universal (Winner, 2009).

Si les tecnologies tenen una dimensió política, doncs, no és només en tant que instruments que poden servir a fins polítics i a una determinada distribució de poder, sinó també en virtut del seu caràcter mediador, que co-determina el marc sociopolític de la vida humana. Amb això es posa de relleu una dimensió de les relacions humà-tecnologia que passa generalment desapercebuda

en el marc de reflexió ètica actual sobre el desplegament tecnològic, i que té una importància normativa decisiva: la dimensió estructural de les relacions humà-tecnologia. Més enllà del nivell 'interpersonal' d'aquestes, en què els humans es relacionen amb les tecnologies en termes d'ús o d'interacció directa amb els artefactes –és a dir, com a usuaris–, les relacions entre humans i tecnologies es donen també en termes estructurals: ens relacionem amb les tecnologies més indirectament en tant que subjectes en estructures socials tecnològicament *mediades*.

D'acord amb el que s'ha exposat fins ara, tant en qualitat d'artefacte com d'activitat, la tecnologia té una dimensió instrumental que, tanmateix, no implica entendre-la com a axiològicament neutral. Tot al contrari: no només la instrumentalitat de la tecnologia respon a un determinat sistema de valors i fins, sinó que, contra el que pressuposa la metafísica humanista, els instruments participen activament en la configuració de la nostra agència i condició moral i, en última instància, en l'estructuració sociopolítica de la vida humana. Així doncs, l'activitat tecnològica tampoc s'explica com a producció de mitjans per a certs fins: es tracta, més fonamentalment, d'una activitat de disseny de mediacions tecnològiques i, per tant, de marcs per a modes de vida.

Des d'aquesta caracterització de la tecnologia, suficientment responsiva a la dimensió moral i política dels artefactes i de l'activitat de desenvolupament dels mateixos, s'esclareix ja el sentit en què la tecnologia concerneix a l'ètica, així com el tipus d'ètica de què es tracta. Aquestes eren les dues primeres qüestions fonamentals per a articular correctament la reflexió ètica entorn la tecnologia.

La tecnologia interpel·la l'ètica no només amb motiu de la seva condició instrumental, que la fa susceptible de ser avaluada moralment en termes d'ús i, per tant, d'efectes i finalitats d'aquest; és a dir, pel fet de ser un mitjà (tant en termes artefactuals com d'activitat) per als fins i interessos humans. La tecnologia interpel·la l'ètica, també, i més fonamentalment, pel fet que, en el seu ús, els artefactes tecnològics *medien* les nostres experiències i accions, configurant així la nostra circumstància moral tant individual com social. Per consegüent, l'ètica de la tecnologia ha d'ocupar-se de la deliberació critico-racional sobre qüestions vinculades al que seria la funcionalitat de la tecnologia (usos, fins, interessos i efectes), però ha d'atendre també el caràcter mediador dels artefactes i les seves implicacions per a les pràctiques en què s'insereixen, així com per a l'estructuració sociopolítica de la vida humana.

D'entrada, doncs, el tipus d'ètica que demana la tecnologia dista molt de ser una reflexió normativa de caràcter extern, centrada en determinar els usos i riscos moralment acceptables o no de les tecnologies[23]. Es tracta d'una ètica que ha d'atendre les pràctiques i societats que la tecnologia contribueix a configurar, i que, en aquest sentit, pot entendre's com una ètica de la concepció i el disseny tecnològic[24]. Per consegüent, el terreny de consideració normativa no concerneix només a la dimensió 'interpersonal' de les relacions humà-tecnologia –és a dir, aquella vinculada a la interacció dels usuaris amb els artefactes en el context immediat d'ús–,

[23] Com s'evidenciarà en el Capítol 3 amb el cas de la robòtica social assistencial, aquests riscos s'associen al funcionament dels artefactes tecnològics i s'entenen sobretot en termes d'impactes en el benestar (individual) dels usuaris.

[24] Per això mateix, com s'insistirà més endavant, l'ètica de l'enginyeria no pot reduir-se a una mera ètica professional, sinó que ha d'atendre a les qüestions sobre el desplegament tecnològic des d'un punt de vista èticament substantiu.

sinó també a la dimensió estructural de les relacions humà-tecnologia. Aquesta és una precisió que cal subratllar, ja que, com es mostrarà en el Capítol 3, el tipus d'ètica de la tecnologia predominant redueix l'esfera de consideració ètica als problemes que sorgeixen en la interacció diàdica humà-tecnologia, negligint greument aquells que tenen a veure amb el rol de les tecnologies en la configuració de les condicions i estructures sociopolítiques de la vida humana. El tipus d'ètica que reclama la tecnologia no pot entendre's com una ètica desvinculada de la política.

1.3. L'estatut de l'ètica en relació a la tecnologia

Arribats aquí, queda pendent clarificar l'estatut de l'ètica de la tecnologia, és a dir, la seva articulació disciplinària. Es tracta d'una ètica aplicada? Aquesta és la manera com s'acostuma a definir contemporàniament (Franssen, Lokhorst and van de Poel, 2023) (Gordon and Nyholm, 2023) (Sætra and Danaher, 2022). D'entrada, podria semblar pertinent entendre l'ètica de la tecnologia com a ètica aplicada: si l'ètica aplicada s'ocupa de problemes específics que apareixen en les diferents esferes d'activitat humana, llavors la tecnologia, en la seva qualitat d'activitat, sembla delimitar-se com un dels camps d'aquesta modalitat de l'ètica; un camp constituït, al seu torn, per un conjunt de subdominis ètics específics corresponents a les diferents activitats tecnològiques (robòtica, computació, IA, ciència de dades, etc.).

Tanmateix, convé examinar la qüestió amb rigor. I és que, lluny de ser una disquisició merament conceptual, determinar si es pot parlar amb propietat d'ètica aplicada per al cas de l'ètica de la tecnologia –i els seus subdominis d'activitat– és una qüestió d'alta importància pràctica. En primer lloc, perquè ajuda a dissipar la manca de claredat i, fins i tot, equívoc generalitzat sobre el procedir de l'ètica de la tecnologia, que es troba latent en l'aproximació ètica actual i en minva la seva força normativa. En segon lloc, perquè contribueix a examinar críticament la tendència de l'ètica de la tecnologia contemporània a la proliferació d'ètiques regionals per a cada tipus de tecnologia (Sætra and Danaher, 2022), que es vincula a una ambigüitat de base sobre l'articulació disciplinària d'aquesta branca de l'ètica.

És, doncs, l'ètica de la tecnologia –i de les tecnologies[25], més concretament– una ètica aplicada? Com s'ha clarificat, la tecnologia té, en última instància, una finalitat externa a si mateixa, en el sentit que es tracta d'una activitat que s'orienta a fins vinculats a altres activitats humanes. En terminologia de MacIntyre, el bé intern de la tecnologia es podria definir en termes de subordinació teleològica. Aquesta condició instrumental de la tecnologia té implicacions decisives per a la respectiva conjugació de l'ètica, que caldrà entendre com una ètica (parcialment)[26] aplicada *subsidiària* de les activitats humanes a què pretén servir la tecnologia. L'ètica de la tecnologia no és, pròpiament, una ètica aplicada, si per això s'entén que el terreny d'aplicació és l'activitat o camp tecnològic en qüestió (i, derivadament, els seus artefactes): precisament per la dimensió instrumental de la tecnologia, la reflexió ètica sobre aquesta no es pot desvincular de les finalitats i valors de les pràctiques per a les quals es conceben i desenvolupen els artefactes. Per tant, si bé és cert que l'ètica de la tecnologia s'ha d'articular

[25] D'ara en endavant, 'ètica de les tecnologies' s'usarà per designar el conjunt de subdominis de l'ètica de la tecnologia d'acord amb els diferents camps tecnològics, i que inclouria l'ètica de la robòtica, l'ètica de la IA, l'ètica de la computació, l'ètica de les dades, etc.

[26] L'adjectivació de l'ètica de la tecnologia com una ètica "parcialment" aplicada s'avança aquí a efectes de disposar d'una definició completa, si bé la raó per fer-ho s'aclarirà més endavant.

com una ètica aplicada –i ha de procedir contextualitzant la reflexió en el marc de les finalitats i valors definitoris de l'activitat tecnològica–, ens trobem amb què, per la seva naturalesa instrumental, aquestes finalitats i valors seran, primàriament, els propis de la pràctica a la que la tecnologia s'orienta com a mitjà.

D'aquí que no sigui rigorós parlar de l'ètica de les tecnologies com a ètica aplicada, si per això s'entén que la reflexió concerneix privativament als camps d'activitat tecnològica i els seus artefactes de per si, sense emmarcar-los en les coordenades teleològiques i axiològiques particulars de cadascun de les pràctiques humanes a què serveixen. Tanmateix, aquest és un dels errors de l'ètica de les tecnologies contemporània, en què la reflexió s'articula principalment entorn el mitjà, en lloc de posar al centre les activitats a què aquest s'orienta, com es mostrarà en el Capítol 3 a través del cas de la robòtica social assistencial (Pareto Boada, Román Maestre and Torras, 2021). L'actual proliferació d'ètiques específiques per a cadascun dels diferents camps tecnològics es podria interpretar, en part, com una manifestació d'aquesta tendència. En sintonia amb el gir empíric de la filosofia de la tecnologia contemporània, l'ètica de la tecnologia s'està ramificant en ètiques correlatives als diferents tipus de tecnologia.

Aquest fenomen ha estat molt oportunament problematitzat per Sætra i Danaher (2022), doncs, en efecte, no procedeix crear una ètica per a cada tecnologia. Segons aquests autors, la proliferació d'ètiques de les tecnologies és innecessària i injustificada, ja que, per norma general, els diferents camps tecnològics no plantegen problemes ètics propis ni nous que donin raó de tal fragmentació disciplinària. Els problemes vinculats a aquestes tecnologies poden ser per tant adreçats des d'una ètica de la tecnologia de gènere, i no pas d'espècie. Alhora, argüeixen que la proliferació és també contraproduent perquè compartimenta la reflexió, afectant negativament la qualitat i eficiència de la tasca de l'ètica en relació a la tecnologia.

Ara bé, més enllà de les raons adduïdes per Sætra and Danaher (2022), la improcedència de la proliferació de subètiques de la tecnologia s'hauria d'avaluar, més fonamentalment, en base a un altre motiu; a saber: el fet que no s'escau amb l'estatut disciplinari de l'ètica de la tecnologia, propi d'una ètica aplicada *subsidiària*. En conformitat amb la dimensió instrumental de la tecnologia, la delimitació de dominis de l'ètica de la tecnologia no hauria de respondre tant a les diferents tecnologies com a les activitats humanes a què aquestes pretenen servir. En aquest sentit, es podria afirmar que la transició històrica que ha fet l'ètica des de la Tecnologia a les tecnologies ha de completar-se amb un gir a les pràctiques humanes tecnològicament 'mediades'.

Per tant, si bé la negativa de Sætra i Danaher (2022) a la qüestió de si cal o no una ètica per a cada tecnologia és essencialment encertada[27], resulta parcialment erràtica pel que fa als motius, doncs la seva argumentació desatén el més fonamental: que, per definició, l'ètica de les tecnologies demana d'una ètica com a hermenèutica crítica de les activitats humanes a les quals aquesta ha de servir i que, per tant, la demarcació de dominis d'estudi no pot respondre pròpiament als diferents camps tecnològics. Atès que els autors parteixen d'una explícita

[27] Això no implica que desestimar una categorització laxa de l'ètica de la tecnologia en subètiques específiques com a manera de fer referència als diferents exercicis d'articulació de l'ètica en el si dels camps d'activitat tecnològica, i que es podria entendre com l'equivalent a una ètica de l'enginyeria entesa no només com a ètica professional sinó també com una ètica de la tecnologia que concerneix als enginyers com a proveïdors d'eines per a pràctiques humanes concretes.

conceptualització de l'ètica de la tecnologia (i les seves derivades) com a "ètica aplicada" (Sætra and Danaher, 2022, p.3), aquest descuit només pot explicar-se com a símptoma de la mala comprensió i ambigüitat generalitzada sobre el procedir d'aquesta modalitat d'ètica i, més particularment, sobre el tipus d'aplicació subsidiària que demana la tecnologia, atesa la seva dimensió instrumental (no neutral).

Reprenent la pregunta per l'estatut de l'ètica de la tecnologia es conclou, doncs, que la seva articulació disciplinària s'ha d'entendre, en primer lloc, com la pròpia d'una ètica aplicada subsidiària a les pràctiques humanes a què s'orienten i serveixen les tecnologies. No obstant, la resposta és encara incompleta: cal precisar el caràcter parcial d'aquesta naturalesa "aplicada" de l'ètica de la tecnologia. Atesa la dimensió moral i política de la tecnologia, l'ètica de la tecnologia és, més concretament, una ètica *parcialment* aplicada: més enllà de reflexionar normativament sobre les tecnologies a la llum del marc teleològic i axiològic definitori de les pràctiques a què serveix (nivell *meso*), s'aproxima a la tecnologia des d'una perspectiva crítica.

'Crítica', en dos sentits fonamentals. Primer, en el sentit que es tracta d'una ètica que posa al centre d'atenció el tipus d'estructuració sociopolítica de la vida humana que les tecnologies contribueixen a configurar (nivell *macro*), escapant així del conservadorisme a què s'arrisca l'ètica aplicada, i que ha estat problematitzat com a dilema estructural d'aquesta disciplina (Kettner, 2003) –la qual, en la seva comesa pragmàtica de donar resposta a problemes concrets, parteix de l' *status quo* i condicions socials preestablertes–. Com s'ha vist, l'ètica de la tecnologia no és una ètica desvinculada de la política, i per consegüent sobrepassa l'exercici d'ètica aplicada. En un segon sentit, perquè es tracta d'una ètica que delibera no només sobre la tecnologia com a mitjà, sinó primàriament sobre els fins de la tecnologia, el que equival a deliberar sobre formes de vida. I és que, parafrasejant Ortega (2004), com a activitat d'un ser que és essencialment projecte (l'humà), la tecnologia està al servei d'un determinat model de vida i, per tant, d'un particular ideal de benestar –de vida bona, *a*-propiada–. Una reflexió ètica sobre la tecnologia implica, en últim terme, un exercici d'imaginació crítica que recupera la pregunta per la causa final de la tecnologia i, per tant, la pregunta pel programa vital al servei de la qual la posem.

En definitiva, l'ètica de la tecnologia és més comprehensiva que una ètica estrictament aplicada, i manté el seu caràcter d'ètica filosòfica substantivament compromesa amb la pregunta per la vida bona, amb la configuració de les nostres formes de vida. Per tant, lluny de reduir-se a una mera demarcació de límits *ex facto*, s'ocupa d'orientar i acompanyar l'activitat tecnològica en vistes al tipus de subjectivitats i societats tecnològicament *mediades* que podem voler legítimament. Es tracta d'una ètica que, en terminologia contemporània, es podria definir com a ètica positiva, si bé l'ètica és sempre, per naturalesa, una activitat de creació de sentit, un afany constructiu-emancipador, més que no pas un exercici d'avaluació moral reactiva.

2. L'abordatge ètic de la robòtica social

De les consideracions anteriors se'n conclou quina és l'articulació de l'ètica en relació a la robòtica social en particular; això és, com abordar des de l'ètica les qüestions normativament rellevants d'aquesta tecnologia, entesa en la seva doble significació d'activitat i artefacte.

D'entrada, l'ètica de la robòtica social es configura, fonamentalment, com una triangulació disciplinar entre ètica aplicada, filosofia política i filosofia de la tecnologia.

En primer lloc, la identificació i anàlisi ètic de les qüestions normativament rellevants que planteja la robòtica social s'ha de fer sempre de forma contextualitzada, atenent al marc de finalitats i valors definitoris de les activitats humanes a què serveixen els robots. Atesa la dimensió instrumental de la tecnologia, l'ètica de la robòtica ha de procedir, en part, com una ètica aplicada subsidiària d'aquestes activitats, emmarcant la reflexió en el seu context teleològic i axiològic. Això significa que l'abordatge ètic de la robòtica social s'ha de fer per camps, però no primàriament tecnològics, sinó relatius a les pràctiques humanes per a les quals es conceben i on s'implementen els robots. Al seu torn, això situa com a part de la reflexió ètica l'anàlisi crític sobre els béns interns i valors d'aquestes pràctiques, i l'objectiu dels robots socials en relació amb elles.

En segon lloc, la reflexió ètica sobre els problemes plantejats per la robòtica social no pot circumscriure's a l'esfera de la interacció diàdica humà-robot. Donada la dimensió política de la tecnologia, l'ètica de la robòtica social ha d'atendre també les relacions humà-tecnologia a nivell estructural, identificant i analitzant les qüestions normativament rellevants que planteja aquesta tecnologia des d'una perspectiva *macro* de justícia. En aquest sentit, l'ètica de la robòtica social ha de desenvolupar la seva tasca de fonamentació critico-racional a partir de recursos conceptuals de la filosofia política.

En tercer lloc, l'ètica de la robòtica social ha d'articular-se en vinculació a la filosofia de la tecnologia. En tant que s'ocupa de reflexionar sobre la tecnologia des de diferents àmbits de coneixement filosòfic (epistemologia, antropologia filosòfica, filosofia política, filosofia del llenguatge), aquesta és una disciplina matriu per a l'ètica de la robòtica social, doncs proporciona el substrat filosòfic per a pensar la tecnologia des de l'ètica. Així ho demostra, per exemple, la contribució que la postfenomenologia suposa per a la reflexió ètica sobre les tecnologies: al posar de relleu el fenomen de la mediació tecnològica de la moralitat, aquesta corrent filosòfica permet dilucidar el sentit en què la tecnologia concerneix l'ètica i el tipus d'ètica de què es tracta.

D'altra banda, les consideracions exposades al llarg d'aquest Capítol també permeten clarificar la relació entre l'anomenada ètica de l'enginyeria i l'ètica de la tecnologia. Tradicionalment, la primera s'entén com una ètica de la professió, centrada en la pràctica laboral, i articulada des d'una comprensió bastant reduïda de responsabilitat per les implicacions del desplegament tecnològic per a la vida humana (Franssen, Lokhorst and van de Poel, 2023). L'ètica de la tecnologia s'entendria, més genèricament, com aquella que s'ocupa dels problemes ètics plantejats per les tecnologies (Sætra and Danaher, 2022). Deixant de banda la confusió contemporània respecte a l'articulació disciplinària de l'ètica de la tecnologia abordada anteriorment, és interessant subratllar la limitació d'aquesta demarcació actual entre l'ètica de l'enginyeria com a una ètica professional desvinculada d'una reflexió ètica sobre les implicacions del desplegament tecnològic per als diferents nivells de la vida humana. En la seva activitat professional els enginyers haurien d'integrar part d'aquesta reflexió ètica més substantiva sobre les tecnologies, atesa la seva dimensió instrumental i el seu caràcter mediador.

Referències

Arras, J. D. (1990) 'Review: Common Law Morality', *The Hastings Center Report*, 20(4), pp. 35–

37.

Bayertz, K. (2003) 'La moral como construcción. Una autorreflexión sobre la ética aplicada', in Cortina, A. and García Marzá, D. (eds) *Razón Pública y éticas aplicadas: los caminos de la razón práctica en una sociedad pluralista*. Tecnos, pp. 47–70.

Brey, P. (2010) 'Philosophy of Technology after the Empirical Turn', *Techné*, 14(1), pp. 36–48.

Camps, V. (2017) *Breve historia de la ética*. Barcelona: RBA Libros.

Coeckelbergh, M. (2018) 'Technology and the good society: A polemical essay on social ontology, political principles, and responsibility for technology', *Technology in Society*. Elsevier Ltd, 52, pp. 4–9. doi: 10.1016/j.techsoc.2016.12.002.

Coeckelbergh, M. (2022) *The Political Philosophy of AI* . Polity Press.

Cortina, A. (1996) 'El estatuto de la ética aplicada. Hermenéutica crítica de las actividades humanas', *Isegoría*, 13, pp. 119–134.

Cortina, A. (2003) 'El quehacer público de las éticas aplicadas: ética cívica transnacional', in Cortina, A. and García-Marzá, D. (eds) *Razón pública y éticas aplicadas. Los caminos de la razón práctica en una sociedad pluralista*. Tecnos, pp. 13–44.

Cortina, A. (2007) *Ética mínima. Introducción a la filosofía práctica*. Tecnos. Madrid.

Cortina, A. and Martínez, E. (2001) *Ética*. 3a Ed. Akal.

Esquirol, J. M. (2011) *Los filósofos contemporáneos y la técnica. De Ortega a Sloterdijk*. Barcelona: Editorial Gedisa.

Feenberg, A. (2010) *Between Reason and Experience: Essays in Technology and Modernity*. Cambridge: MIT Press.

Feenberg, A. (2018) 'What Is Philosophy of Technology?', in Beira, E. and Feenberg, A. (eds) *Tecnology, Modernity, and Democracy. Essays by Andrew Feenberg*. Rowman & Littlefield International.

Franssen, M., Lokhorst, G.-J. and van de Poel, I. (2023) 'Philosophy of Technology', *The Stanford Encyclopedia of Philosophy (Spring 2023 Edition)*. Edward N. Zalta & Uri Nodelman (eds.). Available at: https://plato.stanford.edu/archives/spr2023/entries/technology/.

González García, M. I. and Fernández-Jimeno, N. (2022) 'Introducción. La filosofía de la tecnología y sus identidades múltiples. Una mirada desde España', *Azafea. Revista de Filosofía. Monográfico. Cuestiones actuales en Filosofía de la Tecnología*, 24, pp. 7–19.

Gordon, J.-S. and Nyholm, S. (2023) 'Ethics of Artificial Intelligence', *The Internet Encyclopedia of Philosophy*. Available at: https://iep.utm.edu/ethics-of-artificial-intelligence/.

Heidegger, M. (2009) 'The Question Concerning Technology', in Kaplan, D. M. (ed.) *Readings in the Philosophy of Technology*. 2nd edn. Rowman & Littlefield Publishers, Inc., pp. 9–24.

Hui, Y. (2020) *Fragmentar el futuro: ensayos sobre tecnodiversidad*. Buenos Aires: Caja Negra.

Ihde, D. (1990) *Technology and the Lifeworld*. Indiana University Press.

Ihde, D. (2015) *Postfenomenología y Tecnociencia.Conferencias en la Universidad de Pekín*. Plataforma Editorial Sello.

Kettner, M. (2003) 'Tres Dilemas Estructurales de la Ética Aplicada', in Cortina, A. and García-Marzá, D. (eds) *Razón pública y éticas aplicadas. Los caminos de la razón práctica en una*

sociedad pluralista. Tecnos, pp. 145–158.

Kudina, O. (2019) *The technological mediation of morality: value dynamism, and the complex interaction between ethics and technology*. PhD Thesis, University of Twente.

Kudina, O. and Verbeek, P. P. (2019) 'Ethics from Within: Google Glass, the Collingridge Dilemma, and the Mediated Value of Privacy', *Science Technology and Human Values*, 44(2), pp. 291–314. doi: 10.1177/0162243918793711.

Latour, B. (1993) *We Have Never Been Modern*. Cambridge: Harvard University Press.

Latour, B. (1994) 'On Technical Mediation: Philosophy, Sociology, Genealogy', *Common Knowledge*, Fall V3(2), pp. 29–64.

López Aranguren, J. L. (1991) *De ética y de moral*. Barcelona: Círculo de Lectores.

López Aranguren, J. L. (2005) *Ética*. Madrid: Alianza.

MacIntyre, A. (1984) 'Does applied ethics rest on a mistake?', *The Monist*, 67(4), pp. 498–513. doi: http://www.jstor.org/stable/27902885.

MacIntyre, A. (2019) *Tras la virtud*. Barcelona: Austral.

Mitcham, C. (1994) *Thinking Through Technology. The Path between Engineering and Philosophy*. The University of Chicago Press.

Ortega y Gasset, J. (2004) *Meditación de la técnica y otros ensayos sobre ciencia y filosofía*. 8ª. Revista de Occidente en Alianza Editorial.

Pareto Boada, J., Román Maestre, B. and Torras, C. (2021) 'The ethical issues of social assistive robotics: A critical literature review', *Technology in Society*, 67. doi: 10.1016/j.techsoc.2021.101726.

Pitt, J. (2014) 'Guns Don't Kill, People Kill; Values in and/or around Technologies', in *The Moral Status of Technical Artifacts*. Springer, pp. 89–101.

Ricoeur, P. (2008) *Lo justo 2. Estudios, lecturas y ejercicios de ética aplicada*. Madrid: Trotta.

Román Maestre, B. (2016) *Ética de los servicios sociales*. Herder.

Rosenberger, R. (2017) *Callous Objects: Designs Against the Homeless*. University of Minnesota Press.

Rosenberger, R. and Verbeek, P.-P. (2015) 'A Field Guide to Postphenomenology', in Rosenberger, R. and Verbeek, P.-P. (eds) *Postphenomenological Investigations: Essays on Human-Technology Relations*. Lexington Books.

Ruiz Trujillo, P. (2020) *Ética de las nanotecnologías*. Herder.

Sætra, H. S. and Danaher, J. (2022) 'To Each Technology Its Own Ethics : The Problem of Ethical Proliferation', *Philosophy & Technology*. Springer Netherlands, 35(93), pp. 1–26. doi: 10.1007/s13347-022-00591-7.

Tzafestas, S. G. (2016) *Roboethics. A Navigating Overview*. Springer.

Verbeek, P.-P. (2005) *What Things Do: Philosophical reflections on technology, agency, and design*. The Pennsylvania State University Press.

Verbeek, P.-P. (2009) 'Let's Make Things Better: A Reply to My Readers', *Human Studies*, 32(2), pp. 251–261. doi: 10.1007/s10746-009-9118-0.

Verbeek, P.-P. (2011) *Moralizing Technology: Understanding and Designing the Morality of Things*. The University of Chicago Press.

Verbeek, P. P. (2008) 'Obstetric ultrasound and the technological mediation of morality: A postphenomenological analysis', *Human Studies*, 31(1), pp. 11–26. doi: 10.1007/s10746-007-9079-0.

Verbeek, P. P. (2020) 'Politicizing Postphenomenology', in Miller, G. and Shew, A. (eds) *Reimagining Philosophy and Technology, Reinventing Ihde*. Springer, pp. 141–155. doi: 10.1007/978-3-030-35967-6_9.

Winner, L. (2009) 'Do Artifacts Have Politics?', in Kaplan, D. M. (ed.) *Readings in the Philosophy of Technology*. 2nd edn. Rowman & Littlefield Publishers, Inc., pp. 251–263.

CAPÍTOL 3. COMPENDI DE PUBLICACIONS

1. Prolegómenos a una ética para la robótica social*

Prolegomena to an Ethics for Social Robotics

Júlia Pareto Boada

Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Llorens i Artigas 4-6, 08028 Barcelona, Spain
jpareto@iri.upc.edu
ORCID iD: 0000-0003-4879-8800

Resumen: La robótica social presenta un elevado potencial disruptivo, al expandir el ámbito de aplicación de la tecnología inteligente a contextos prácticos de naturaleza relacional. Por su capacidad de interactuar con las personas "intersubjetivamente", los robots sociales pueden asumir nuevos roles en nuestras actividades cotidianas, multiplicando las implicaciones éticas de la robótica inteligente. En este artículo ofrecemos algunas consideraciones preliminares para la reflexión ética sobre la robótica social, para clarificar cómo orientar acertadamente el pensar crítico-normativo en esta ardua tarea. Defendemos la 'perspectiva del ser' y sus categorías vinculadas de 'teleología' e 'interés' como aquellas desde las que articular la reflexión. Argumentamos que atender primariamente al 'ser' de los robots, antes que a su 'hacer', permite acercarnos correctamente al fenómeno nuclear de toda preocupación ética sobre la robótica social (la externalización de nuestra agencia en estas entidades), evitando los reduccionismos en la mirada ética a que éste puede conducir.

Palabras clave: agencia, ética, inteligencia artificial, interacción humano-robot, robótica social

Abstract: Social robotics has a high disruptive potential, for it expands the field of application of intelligent technology to practical contexts of a relational nature. Due to their capacity to "intersubjectively" interact with people, social robots can take over new roles in our daily activities, multiplying the ethical implications of intelligent robotics. In this paper, we offer some preliminary considerations for the ethical reflection on social robotics, so that to clarify how to correctly orient the critical-normative thinking in this arduous task. We defend the 'being perspective' and its linked categories of 'teleology' and 'interest' as the ones from which to articulate the reflection. We argue that attending primarily to the robots' 'being', before their 'doing', allows us to correctly approach the core phenomenon of all ethical concerns on social robotics (the outsourcing of our agency in these entities), avoiding the reductionisms in the ethical gaze to which the latter may lead.

Keywords: agency, artificial intelligence, ethics, human-robot interaction, social robotics

1. Ante la fuerza disruptiva de la robótica social

El despliegue de sistemas de inteligencia artificial en el ámbito de la robótica ha multiplicado, en las últimas décadas, el potencial disruptivo de este campo tecnocientífico. La "encarnación" de inteligencia artificial en dispositivos robóticos da lugar a "agentes artificiales" capaces de

---

*** The text of this section entirely corresponds to the following publication: Pareto Boada, J. (2021) 'Prolegómenos a una ética para la robótica social', en Jon Rueda (ed.): Tecnologías socialmente disruptivas. *Dilemata, Revista Internacional de Éticas Aplicadas, (34)*, pp. 71–87.**

llevar a cabo, con cierto grado de autonomía tecnológica (Funk y Coeckelbergh, 2020), acciones[28] con objetivos específicos en entornos reales con independencia de control externo. Podemos encomendarles no sólo tareas automatizables, sino conductas que requieren de un previo procesamiento cognitivo de información sobre el entorno y de elecciones a partir de ella, siempre respecto a un fin predeterminado.

Con ello la robótica inteligente altera radicalmente el abanico de tareas y roles que pueden asumir los robots, extendiendo sus posibles ámbitos de aplicación a contextos prácticos que hasta ahora quedaban reservados a la agencia humana. En este sentido hablamos de disrupción, pues la expansión del potencial robótico a tareas que exigen una interacción y adaptabilidad al entorno nos sitúa ante una posible transformación de las actividades humanas, tanto a nivel práctico-estructural como conceptual. A su vez, se ensancha el conjunto de implicaciones éticas de la robótica inteligente y, por tanto, el alcance de la reflexión normativa necesaria para que la introducción de sus artefactos en nuestra existencia se halle en consonancia con los valores morales y derechos humanos. A esto responde, precisamente, la aparición de nuevas disciplinas como la 'Roboética' (Operto y Veruggio, 2008), la 'Ética para Máquinas' (*machine ethics*) y la todavía más incipiente 'Robofilosofía' (Seibt, 2017).

Uno de los grandes focos de disrupción de la robótica inteligente se encuentra en los "robots sociales", destinados a servir en nuestra cotidianidad de forma radicalmente nueva, a saber, como entidades capaces de entrar en nuestra dinámica humana de interacción social, no sólo física sino también semánticamente (Rodogno, 2016). Esto les confiere un cierto estatuto de agentes sociales que parece situarlos en una ontología resbaladiza, por cuanto presentan características de 'casi-sujetos'. Así, los robots sociales, al asumir un rol en prácticas socioculturales, irrumpen en nuestra forma de concebir y organizar la vida en comunidad. A esta clase de sistemas inteligentes autónomos pertenecerían, por ejemplo, los robots personales en el ámbito doméstico para la asistencia en situaciones de vulnerabilidad, los robots para los cuidados en el sector sociosanitario, los robots asistenciales para contextos terapéuticos de rehabilitación cognitiva, o los robots de apoyo a la educación.

La fuerza disruptiva de estas inteligencias artificiales "socialmente situadas" ubica la rama de la robótica social bajo un fuerte escrutinio ético. Esta actividad tecnocientífica y sus productos desatan crecientes preocupaciones sobre un sinfín de cuestiones relativas a la dignidad humana, la autonomía personal, la libertad, la privacidad, la responsabilidad, la devaluación de prácticas humanas como los cuidados, la degeneración de nuestras facultades morales, la justicia distributiva, etc. Así, en la actualidad, la explosión de debates y reflexiones sobre las múltiples

[28] El concepto de 'acción' se ha adoptado comúnmente para hablar de las tareas o funciones de los robots inteligentes, de lo que estos *hacen*. Inspirándonos en la diferenciación señalada por Arendt (2005) entre 'labor', 'trabajo' y 'acción', sería oportuna en la ética para la robótica social la distinción entre la capacidad de obrar (acción) y la capacidad de hacer (labor y trabajo), para diferenciar entre dos tipos de agencia: aquella autónoma y moral, caracterizada por la espontaneidad, la imprevisibilidad y la pluralidad y, por tanto, exclusivamente humana, y aquella agencia no moral, a la que pertenecerían los actos heterónomos de entidades fabricadas, como los robots. Esto conllevaría diferenciar connotativamente entre los términos 'acción/acto/obrar' y 'quehacer/tarea/conducta/comportamiento'. Referirnos con propiedad a la dimensión agente del robot implicaría usar siempre las nociones pertenecientes al segundo grupo, por lo que no hablaríamos de 'actos' o 'acciones' de los robots, sino de 'quehaceres' o 'conductas'. En este trabajo trataremos de mantenernos fieles, a partir de ahora, a esta distinción sustancial.

implicaciones de la robótica social a diferentes niveles (individual, interpersonal, socio-estructural, profesional, etc.) responde al reto colectivo de orientar el desarrollo, implementación y uso de estas aplicaciones de la inteligencia artificial (Nørskov et al., 2020; Coeckelbergh et al., 2018; Seibt et al., 2016).

El objetivo del presente artículo es contribuir a esta deliberación crítico-normativa mediante una serie de consideraciones preliminares a la reflexión ética sobre la robótica social. De ese modo, no nos ocuparemos de discutir un problema concreto de la robótica social, sino que, en un estadio previo, señalaremos algunos rasgos fundamentales de la perspectiva desde la cual dirigir la mirada ética. Se pretende, en definitiva, establecer ciertas bases que ayuden a clarificar la compleja cuestión de cómo pensar la robótica social desde el punto de vista de la ética, asentando algunos fundamentos para poder desvelar los retos relevantes que plantea a nivel ético-normativo. Y es que, como advierte Žižek (2013), para encontrar respuestas debemos formular antes la pregunta correcta.

Con vistas a este propósito, se empezará abordando, en el segundo apartado, el controvertido concepto de "robot social", para acotar, a su vez, la rama de actividad tecnocientífica en la que nos centraremos. En el tercer apartado expondremos la complejidad que supone reflexionar en torno a la robótica social desde una perspectiva ética. En el cuarto apartado, se identificará el fenómeno nuclear de las preocupaciones éticas sobre la robótica social (que designamos como 'externalización de la agencia'), y se advertirá del posible reduccionismo en la mirada ética a que éste puede conducir. De este modo, se revelará lo que consideramos una aproximación desenfocada, articulada primariamente en el 'hacer' del robot. En el quinto apartado se argumentará a favor de una perspectiva ética centrada prioritariamente en la cuestión del 'ser'. Se desvelarán dos categorías (teleología e interés) clave para la reflexión, que parecen haber pasado inadvertidas en el marco de discusión actual sobre la robótica inteligente.

2. Sobre la noción de robot social

Como evidencian Sarrica et al. (2020) en su revisión sistemática de literatura científica y popular, no existe un consenso en la definición de "robot social", si bien hay algunas conceptualizaciones de referencia. Su recopilación de definiciones nos permite concluir que estas difieren según los rasgos en que se ponga el énfasis (función, objetivo, aspecto, capacidades del robot, etc.).

Para los propósitos de este artículo, consideramos pertinente caracterizar los robots sociales como aquellos concebidos para realizar tareas en el marco de prácticas que involucran cierta interacción social, como es el caso de los contextos de cuidados, educación o compañía. Su particularidad consiste en ser "agentes artificiales" que proveen asistencia *por medio de* "interactuar socialmente" con los humanos (aunque no necesariamente en exclusiva con éstos).

En el ámbito tecnocientífico que nos ocupa, dicha interacción se entiende en un sentido muy concreto: que el robot se interrelaciona con nosotros de "forma interpersonal" (Breazeal, Takanishi and Kobayashi, 2008); lo que a veces lleva incluso a calificar los robots sociales como aquellos destinados a relacionarse con las personas a nivel emocional (Campa y Campa, 2016). Con ello básicamente se pretende distinguir el carácter específico de este tipo de bidireccionalidad comunicativa de otras interacciones que se dan entre humanos y robots sin necesidad (o, mejor dicho, sin intención predeterminada) de "relación" psico-emocional.

El rasgo más distintivo de los robots sociales es, pues, su capacidad de interactuar "humanamente", siguiendo patrones básicos de lo que se consideran interacciones intersubjetivas significativas. Ahora bien, es fundamental no perder de vista el marco 'funcionalidad-finalidad' donde se ubica esta capacidad de interacción, pues su ejercicio siempre se orienta a un objetivo concreto. En este sentido, por hacer hincapié en el esquema de finalidades que entra en juego en un robot social, resulta clarificadora la definición de Sheridan (2020), que señala que el objetivo del robot es establecer con la persona una interacción afectiva o de otro modo útil –esto es, siempre provechosa con vistas a un fin.

Esta puntualización permite situarse más rápidamente en las ambigüedades de la noción "robot social". Para empezar, se entiende mejor el sentido en que se usa comúnmente el término "social" en relación a este tipo de inteligencias corporeizadas. A veces, el contraste entre los robots sociales y los industriales puede confundir, conduciendo incluso a una equiparación inexacta entre robot social y robot de servicios. Fundamentalmente, 'social' se dice de un robot por como este artefacto desarrolla sus tareas, es decir, interactuando. Cuál sea la finalidad última de esta interacción es ya otra cuestión, que tiene que ver con su campo de aplicación o el perfil del usuario principal. Con ello se dilucida también la razón de ciertas categorizaciones clásicas en la literatura sobre robótica social. Dado que la interacción humano-robot tiene como finalidad servir a algún objetivo dentro de un determinado contexto relacional, los robots sociales se pueden clasificar en distintos grupos de acuerdo con criterios relativos a esta finalidad.

Uno de ellos sería, por ejemplo, el ámbito de las prácticas en las que asiste el robot (educativo, sociosanitario, de cuidados, asistencial, terapéutico, doméstico, etc.). A esto responde la gran diversidad terminológica (a menudo poco rigurosa) para categorizar robots sociales en distintos (sub)grupos: "robots tutores o profesores", "robots sanitarios", "robots de cuidados" (*care robots*), "robots asistenciales", "robots personales", "robots de compañía", etc.

Otro criterio detectable es lo que podríamos entender como el carácter de la finalidad última de la interacción del robot con el humano. Parece haber cierto empeño en diferenciar los robots sociales en función de si la razón última de su interacción es externa o interna a la misma –es decir, si la interacción es parte necesaria pero no suficiente de la funcionalidad del robot (esto es, la interacción no agota la tarea del robot) o si la interacción y la funcionalidad se equiparan (la interacción es la tarea).

A esta distinción respondería, por ejemplo, la taxonomía de Feil-Seifer y Matarić (2005), en la que el concepto de "robots socialmente asistenciales" (*socially assistive robots* o *SAR*) –más oportunamente redefinidos como "robots sociales asistenciales" en (Payr, 2015)– se introduce para singularizar aquellos robots cuya interacción social es un medio para conseguir desarrollar una tarea de apoyo en un ámbito determinado. Éste sería el caso de un robot que asistiera en el campo de la salud mental ayudando a un usuario a llevar a cabo un ejercicio cognitivo a través de una comunicación verbal y gestual. Un ejemplo de ello es el sistema cognitivo robótico para asistir a pacientes[29] con demencia en un juego de estimulación cognitiva, desarrollado por

[29] Adviértase que este sistema cognitivo robótico se concibe principalmente como una herramienta de apoyo a los terapeutas, por cuanto les ayuda a realizar esta tarea concreta en el marco de sus sesiones con los pacientes. Son estos profesionales quienes se encargan de preprogramar el robot para su función con cada usuario específico, a quien luego el sistema se va adaptando durante el transcurso de la tarea.

Andriella et al. (2020) en el Institut de Robòtica i Informàtica Industrial (CSIC-UPC), y en colaboración con la Fundació ACE. Gracias a su capacidad de percepción, aprendizaje y reacción al comportamiento del usuario, el robot interactúa de forma adaptativa, moldeando autónomamente en cada momento, según el historial de sus acciones y el estado del ejercicio, el nivel y tipo de asistencia a ofrecerle (animando, recomendando o desvelando la solución).

La intención de Feil-Seifer y Matarić (2005) es contraponer esta clase de robots sociales a otros cuya interacción social es la finalidad última, no habiendo así una tarea posterior con respecto a la cual esta es un medio. Esto no impide discernir los motivos por los que se busca esta interacción, lo que da lugar a (sub)clasificarlos en robots de compañía, entretenimiento, apoyo emocional en diferentes ámbitos, etc. Cabe señalar que a los robots cuya funcionalidad es la interacción por la interacción, Feil-Seifer y Matarić (2005) los identifican con los que Fong et al. (2003) habían tipificado de "robots socialmente interactivos" (*socially interactive robots* o *SIR)*, si bien no parece que originalmente esta categoría excluyera los robots cuya función interactiva tuviera como objetivo una tarea externa a la misma. Por el contrario, esta terminología vendría a reforzar la idea de que el rasgo distintivo de los robots sociales es, precisamente, interactuar como *manera de* desarrollar una tarea, independientemente de si ésta se agota o no en la interacción.

Una distinción excesivamente perfilada del carácter de la finalidad última de la interacción tiene implicaciones éticas. Estimamos importante señalar que, en relación a ciertas prácticas de tipo relacional a las que se pretende asistir tecnológicamente, centrar el desarrollo robótico en una visión desmesuradamente fijada en tareas puede suponer un empobrecimiento de la actividad. Por ello, cuando la interacción social se reduce a una forma de llevar cierta labor a buen puerto, limitando la interacción a un objetivo específico y discerniendo la tarea del acto social, se puede reforzar una forma perjudicial de fragmentar actividades humanas de dimensión holística. La robótica social se insiere siempre, en última instancia, en el contexto de cuidados, entendido en un sentido comprehensivo –esto es, no como relativo a un paradigma clínico/médico-rehabilitador, sino a uno social-hermenéutico cultural–. En este sentido, el cuidado no significa paliar déficits de salud, sino que se vincula a una noción integral de salud en que entran en juego la prevención y el bienestar psicoemocional, entre otros; es decir, se trata de una actividad orientada al sustento del "mundo" de las personas (Tronto, 1993). Esto problematiza distinguir demasiado estrictamente entre las finalidades externas e internas de las funcionalidades interactivas para las que se diseña un robot en ámbitos prácticos de esta índole, pues en las tareas pertenecientes a este tipo de prácticas (asistencia física, cognitiva o emocional), 'asistir' y 'cuidar' se interrelacionan, y la finalidad de la interacción difícilmente se puede delimitar con exactitud.

3. ¿Cómo orientar el pensar ético en torno a la robótica social?

En la creciente literatura académica sobre las implicaciones éticas de la robótica social constatamos un desorden abrumador. Existe una gran diversidad de criterios y perspectivas –no siempre explicitados y todavía menos frecuentemente justificados– desde los que se identifican y se toma posicionamiento con respecto a las problemáticas (Vandemeulebroucke, Dierckx de Casterlé and Gastmans, 2018). Esto da lugar a preocupaciones de índole muy

distinta[30]. Sin duda, este embrollo es sintomático de la complejidad que entraña la robótica social como objeto de análisis crítico-reflexivo, es decir, a la hora de ser abordada desde la ética.

3.1. La ética como actividad reflexiva

En general, se da por supuesta la comprensión de lo que significa aproximarse a la robótica social desde una perspectiva ética. Rara vez encontramos una explicación del término "ética" en la literatura crítico-normativa sobre robótica social, y mucho menos una explicitación del sentido de la relación entre ambas actividades.

El desafío reflexivo al que nos emplazan los nuevos horizontes abiertos por la robótica inteligente requiere distinguir la 'ética' de la 'moral'. Ambas tienen que ver con el ejercicio de orientación de la acción humana. Ahora bien, aunque son etimológicamente equivalentes y se usan popularmente de forma indistinta, hay consenso acerca de la necesidad de entender estos dos términos como connotativamente distintos (Ricoeur, 2008). El motivo es poder capturar dos dimensiones de nuestra vida práctica que necesitamos diferenciar: las acciones y las razones. La ética se entiende como una reflexión crítico-racional sobre los fundamentos de la moral, esto es, sobre la validez de las normas y valores de nuestros hábitos y costumbres. De ahí la conocida contraposición entre la moral como "vivida" y la ética como "pensada" (López Aranguren, 1994). La ética se ocupa de la legitimidad, de deliberar acerca de la necesidad de mantener, abandonar o (re)generar las distintas morales –que, en definitiva, son prácticas relativas a determinados contextos socioculturales e históricos y, por tanto, revisables–. La moral se entiende como adhesión a valores, reglas, a partir de los cuales las acciones se estiman como 'buenas' o 'malas'. Por esto la ética se define como una metamoral, al ser una "reflexión de segundo grado sobre las normas" (Ricoeur, 2008, 48). De este modo, y siguiendo a (Cortina, 2007), podemos diferenciar la moral y la ética por su forma inmediata o mediata de orientar la esfera de las acciones humanas respectivamente. Pues mientras la moral responde a la pregunta "¿qué debo hacer?" –indicando una acción u omisión–, la ética contesta la cuestión de "¿por qué debo [hacerlo o no]?" (Cortina, 2007, 62) –ofreciendo, por tanto, argumentos.

Es a la ética así entendida, y no a la moral, a la que corresponde afrontar el reto de orientar el despliegue de la robótica social. Ante este nuevo escenario, lo que necesitamos es poder fundamentar las acciones a emprender en relación al desarrollo, implementación y uso de estos sistemas inteligentes autónomos en base a unos valores crítica y explícitamente analizados y revisados, aprovechando el potencial disruptivo de la innovación para poner entre paréntesis el 'automatismo moral' que suele regir nuestra cotidianidad. Esto es una tarea colectiva que requiere de interdisciplinariedad, y no sólo se opone a las posiciones extremas de rechazo o

[30] Así lo muestran los resultados del trabajo de revisión bibliográfica (en curso) sobre las preocupaciones éticas planteadas en relación a la robótica social asistencial, llevado a cabo en coautoría con las Dras. Begoña Román y Carme Torras. En la fecha de publicación del presente artículo, susodicho trabajo se encuentra en proceso de redacción para su edición.
El proceso de revisión se ha centrado en la literatura de la base de datos Scopus comprendida entre los años 2006 y 2020. Para la búsqueda, se han usado cuatro entradas terminológicas: "ethics" AND "assistive robot*" OR "care robot*" OR "social robot*" OR "human-robot interaction". Con ello se han identificado una gran diversidad de cuestiones éticas asociadas a este campo de la robótica inteligente. Para cada entrada, las problemáticas se han compilado y categorizado en tablas que facilitan su visualización, en base a las cuales se ofrece un análisis cuantitativo-descriptivo del estado de la cuestión, complementado con un análisis cualitativo-crítico posterior. Estas tablas se incluirán en este próximo artículo previsto.

acogida acrítica de los avances tecnológicos, sino que contrasta con ellas por tomar posición activa en la (re)configuración de la realidad social.

Cabe mencionar que, desde un punto de vista ético, hay distintas perspectivas desde las que se puede fundamentar el porqué de una acción, según el carácter de las razones desde las que se examina su corrección o incorrección. Principalmente, la reflexión se puede articular a partir de principios, consecuencias o virtudes, lo que explica la tradicional clasificación entre éticas deontológicas, teleológicas y de la virtud (Camps, 2017). En el panorama actual de deliberación ética sobre la robótica social hay una considerable heterogeneidad a nivel de perspectivas. La detección y exposición de los problemas se suele argumentar desde uno de estos tres criterios o ángulos de valoración. Es importante recordar que una actividad ética debería ser inclusiva, fundamentando la acción no solo en principios o consecuencias, sino integrando armónicamente a ambos en la consideración (Weber, 2012), y siempre en vista del tipo de sociedades que queremos. Todo ello sin desdeñar la atención al contexto, cuya relevancia ha sido bien advertida por la ética del cuidado (Busquets Surribas, 2019): pues no sólo hay valores inherentes a este que deberán introducirse en la ponderación (como cuando hablamos de contextos profesionales), sino que también las consecuencias de una misma acción pueden variar según las circunstancias y sus afectados.

3.2. Las complejidades de una deliberación ética

Hay ciertos elementos que hacen notablemente complicado reflexionar sobre la robótica social desde una perspectiva ética y dirigir el pensamiento en este terreno de forma apropiada y fructífera.

En primer lugar, existe una considerable diversidad de esferas de acción implicadas por la actividad de la robótica social y respecto a las cuales, por tanto, enfocar el pensar. Dado que la ética es una actividad de reflexión centrada en la argumentación del porqué de cursos de acción, esto supone una primera dificultad de orientación, pues, con vistas a este fin, hay una multiplicidad de ámbitos de acción a barajar: (1) la robótica social como actividad tecnocientífica con diversos sujetos implicados, pluralidad de intereses y valores (Echeverría, 2003); (2) las decisiones concretas de los ingenieros en cada una de las fases constitutivas de la actividad (diseño, desarrollo y experimentación de los productos); (3) las acciones relativas a la implementación de los sistemas inteligentes resultantes (adopción institucional, integración de los artefactos en ámbitos profesionales); (4) las acciones de los usuarios; etc. Esto comporta, a su vez, una variedad de sujetos a quienes conciernen las reflexiones y determinaciones que se derivan de la reflexión ético-normativa. Este intrincado entramado de dimensiones es al que quiere dar respuesta la 'Roboética' en su sentido primario (Veruggio y Abney, 2014), esto es, como disciplina centrada en el actuar humano en las diferentes fases de despliegue de la robótica y sus productos y, por tanto, ocupada con los fundamentos de acción de sujetos pertenecientes a campos de actividad muy diversos.

Parece que a ello se añadiría, además, otro ámbito para la consideración ética: el hacer de los robots sociales. Dado su estatuto de sistemas inteligentes con autonomía para (inter)actuar con el entorno para el logro de fines –esto es, para percibir, procesar información y obrar en consecuencia, adaptándose y aprendiendo en el curso de interacción–, estos robots plantean cuestiones relativas a su comportamiento y capacidad de sopesar los cursos del quehacer en

contextos humanos que requieren de cierta "sensibilidad para consideraciones morales" (Wallach y Allen, 2009, 34). Esta es la línea de reflexión de la 'Ética para máquinas', la segunda de las ramas de la 'Roboética'. Volveremos a ello más adelante.

En segundo lugar, la naturaleza tecnológica de la robótica social dificulta el abordaje ético de la misma, al precisar de un 'pensar ético por esferas (conjugadas)', que debe atravesar varias capas de análisis. El carácter instrumental de la tecnología requiere contextualizar la reflexión ética de acuerdo con el campo de aplicación de la robótica. Dado que la finalidad de la robótica social es, en última instancia, externa a la misma –consistente en servir a los fines del contexto práctico en que sus productos se introducen como medios de apoyo–, la reflexión ética no puede desvincularse de la esfera de actividad humana concreta para la que se desarrollan los artefactos. Esto requiere filtrar el pensar ético a la luz de un determinado campo de aplicación de la robótica social (asistencial, educativo, sociosanitario, terapéutico, etc.), e interpretar los retos éticos que comporta en el marco de valores y finalidades propios de estos contextos –sin olvidar el universo más general de principios y valores de la ética cívica donde se legitiman.

Por consiguiente, una reflexión ética en torno a la robótica social debería proceder, al menos en parte, como una ética aplicada, en el sentido de enfocarse a cada uno de los diferentes sectores para los que se desarrolla esta actividad tecnocientífica. Esto es importante, pues en realidad no es riguroso hablar de una 'ética aplicada a la robótica social', si por esto se entiende que el terreno de reflexión es exclusivamente relativo a esta actividad y/o sus artefactos sin enmarcarla en las coordenadas de principios, valores y finalidades particulares de cada contexto práctico para el que se conciben y se desarrollan. Por eso preferimos hablar de ética *para* la robótica social. Se trataría de un ejercicio de ética aplicada como hermenéutica crítica de las actividades humanas (Cortina, 1996). No ayuda a resolver problemas, que es a lo que se dedica la ética aplicada (Román Maestre, 2016), plantear las problemáticas de la robótica social sin "situarlas" en una determinada actividad.

En tercer lugar, en el seno de los sistemas inteligentes de la robótica social se da una cierta convergencia entre dos estatutos ontológicos hasta ahora claramente diferenciados, a saber: sujeto y objeto. Esto distorsiona con facilidad el enfoque de la reflexión ética acerca de esta actividad tecnocientífica. Los robots sociales, como entidades con capacidad de hacer como resultado de un proceso cognoscitivo previo, plantean retos éticos no solo en tanto que objetos –ahí surgen cuestiones sobre privacidad, impacto laboral, desigualdad social, robotización de los cuidados, etc.–, sino también en tanto que "sujetos agentes" que desarrollan una tarea dentro del contexto de actividades humanas cotidianas, y lo hacen mediante una interacción pretendidamente "intersubjetiva". Es por esta capacidad de agencia que la 'ética para máquinas' se orienta a garantizar el alineamiento de los sistemas inteligentes autónomos con los valores humanos. Por el momento, adelantaremos que la piedra de toque de una aproximación ética rigurosa a la robótica social consiste en una determinada concepción de esta "agencia" artificial de sus artefactos.

4. Autonomía tecnológica y externalización de la agencia: ¿de qué hablamos?

La creciente autonomía tecnológica de los robots inteligentes abre la posibilidad de transferirles parte de nuestra agencia para la realización de tareas concretas. Es lo que denominaremos 'externalización de la agencia humana', por cuanto tiene que ver con *encomendar* a entidades

tecnológicas determinados quehaceres que requieren de facultades cognitivas e interactivas con el entorno.

Es primordialmente en virtud de este fenómeno que la robótica inteligente, así como su rama de la robótica social, urge a ser pensada desde la ética. El hecho de externalizar parte de nuestra agencia, delegándola o asignándola a los robots, nos sitúa ante un desdibujamiento de los pilares de nuestra vida práctica. Se plantean interrogantes sobre qué roles confiarles a los robots, para qué y las transformaciones que ello supone, tanto en un contexto de prácticas determinado, como en el plano social e individual.

Efectivamente, si examinamos las diferentes preocupaciones éticas planteadas por la robótica social en los debates actuales descubrimos que todas ellas pivotan, ya sea más explícita o implícitamente, sobre el fenómeno de la externalización de la agencia. Piénsese, por ejemplo, en las recurrentes controversias sobre la vulneración de la dignidad humana, como la objetificación de los usuarios de robots sociales asistenciales o el engaño de una relación "intersubjetiva" con robots de compañía (Coeckelbergh, 2012; Noori et al., 2019; Sharkey y Sharkey, 2012; Sparrow y Sparrow, 2006). En la misma línea están las discusiones sobre los límites de la interferencia de un robot social con la autonomía o la privacidad personal (Feil-Seifer y Mataric, 2011), y los debates acerca de la responsabilidad por posibles daños inesperados. O repárese en las inquietudes sobre la denigración de facultades morales humanas que conllevaría tanto la sustitución del profesional por robots en prácticas de cuidado (O'Brolcháin, 2019; Vallor, 2015), como la interacción humano-robot (Cappuccio, Peeters and McDonald, 2020); o la corrupción de prácticas de cuidados a raíz de la robotización de tareas en contextos hermenéuticos. Todas estas problemáticas vienen desencadenadas por el hecho de que nuestra capacidad de realizar tareas específicas sea asumida por esta tipología de agentes artificiales.

Ahora bien, resulta crucial advertir los reduccionismos en la mirada ética a que puede conducir el fenómeno de externalización de la agencia humana. Éste, sin un previo análisis crítico de la perspectiva bajo la cual entenderlo debidamente, parece impelernos a estructurar la aproximación ética a la robótica social de forma desacertada, centrada en el 'hacer' del robot.

En efecto, si los robots sociales son concebidos como entidades a quienes encargar (parte de) nuestras funciones en prácticas cotidianas de naturaleza relacional, fácilmente puede darse una fijación en cuestiones relativas a su dimensión "agente". Dada su capacidad de asumir un rol encomendado, deviene esencial determinar cómo estos robots deberán comportarse e (inter)actuar para que puedan entrar legítimamente en nuestro espacio social y hacerse cargo de susodichas tareas de una determinada manera (apropiada a la nuestra, no a la suya).

Según la premisa de que depende necesariamente de su comportamiento que un robot sea o no beneficioso para la humanidad (Veruggio y Abney, 2014), se ha dirigido la atención a cuestiones propias de la 'ética para máquinas', orientada a dotar (en caso de ser posible) a los robots de competencia moral. Se aduce (Allen, Wallach and Smit, 2006) que esta es el requisito para asegurar que el impacto de su quehacer en entornos inherentemente éticos –no estructurados y, por tanto, complejos e impredecibles– sea positivo, en el sentido de estar alineado con los principios y valores humanos (Cave *et al.*, 2019). A pesar de existir sólidas refutaciones a la idea de que implementar capacidad de razonamiento ético a sistemas

inteligentes sea garantía de resultados sociales positivos (Brundage, 2014), se destinan numerosos esfuerzos a ello (Sharkey, 2020), para que estos puedan sopesar adecuadamente los posibles cursos de su quehacer de acuerdo con ciertos principios, cuya concreción y translación computacional es otro motivo de debate.

Sin duda, las cuestiones relativas al 'hacer' del robot son relevantes para la orientación normativa de la robótica social. No obstante, centrar la reflexión *primariamente* en esta dimensión supone pasar por alto el foco de atención verdaderamente principal y, por tanto, un reduccionismo de la mirada ética. En ello, juega un papel determinante el marco desde el que nos acerquemos al fenómeno de la externalización de la agencia.

Se pueden identificar una serie de aspectos implícitos y constitutivos del modelo de aproximación ética desenfocada a la robótica social, en el que las problemáticas se piensan, se definen y se juzgan prioritariamente a la luz de la cuestión por el 'hacer'. Los introducimos someramente.

El más relevante es una concepción de los robots como agentes artificiales que se insieren en nuestra cotidianidad en calidad de sujetos o "casi-otros". Esta perspectiva es la que Coeckelbergh (2011) considera como perteneciente a una "ontología individualista", bajo la que los robots sociales son vistos como "individuos" –es decir, como un foco bien demarcado y 'cerrado' de agencia–. Desde esta óptica, el fenómeno de externalización de la agencia casi naturalmente induce a una aproximación ética deficiente, que parte de una idea de los robots como "sujetos" diferenciados de los humanos, que entran en nuestros contextos prácticos como "alteridades" que tienen su propia agencia (aunque delimitada, por heterónoma y hetero-referente).

De ahí que los objetivos de la 'ética para máquinas' se presenten como apremiantes y se articule la reflexión ética sobre la robótica social primeramente en torno a cuestiones relativas al estatuto moral de los robots, no sólo como sujetos agentes, sino también pacientes (Coeckelbergh, 2020; Gunkel, 2018).

A la par, todo esto se vincula a una fijación ética en los impactos que los robots tienen por su conducta, y que, además, se limitan al ámbito individual de la vida humana. Pues cabe señalar una forma diádica de entender la relación humano-robot, que articula la reflexión ética sobre este binomio. Se centra entonces la atención en los impactos que el robot social puede tener sobre la persona con quien interactúa en el desarrollo de una tarea, a riesgo de descuidar algo fundamental: que dicho artefacto se introduce en un entramado de por sí relacional, esto es, en una red de interrelaciones entre distintos actores que configuran la totalidad de la actividad (Vallès-Peris, Angulo and Domènech, 2018).

5. Una aproximación ética a la robótica social desde el 'ser': teleología e interés

Aquí defendemos que la cuestión fundamental a cuya luz acercarnos a la robótica social desde una perspectiva ética no es la del 'hacer', sino la del 'ser'. Antes que 'hacer correctamente', un robot social debe 'ser correcto', lo cual tiene que ver con la idea de legitimidad. Preguntarse por el ser implica tener en cuenta la coherencia entre la razón de ser del agente artificial y las finalidades a las que sirve a través de sus funcionalidades; es decir, significa pensar en términos

de adecuación entre acciones, valores y objetivos, entrando todos ellos en la ecuación como objetos de la actividad crítico-reflexiva.

La relevancia de esta dimensión no ha pasado totalmente desapercibida, pues hay autores que reflexionan sobre la pertinencia de otros tipos de tecnología partiendo de aspectos como, por ejemplo, los problemas que pretenden solventar (Baumer y Silberman, 2011) o los valores respaldados por su diseño (Millar, 2015). Sin embargo, consideramos imperativo explicitar la centralidad de esta cuestión, ya que es aquella que debería moldear primariamente la perspectiva desde la que reflexionar éticamente en torno a la robótica social, por lo que respecta a todas las fases de su despliegue.

Desde el prisma del 'ser' se explicitan dos categorías fundamentales para una aproximación ética pertinente a la robótica social: teleología e interés. Ambas promueven una mirada adecuada de los robots sociales como artefactos que, como productos fabricados por (y para) la actividad humana (Johnson G., 2011), nacen en un marco de finalidades concretas que responden a intereses determinados –que están siempre vinculados a un contexto político, sociocultural e histórico específico y que también deben ser tomados en consideración en la reflexión ética.

Con ello se hace manifiesto que la agencia de estos artefactos inteligentes responde a un propósito último, que le viene impuesto, asignado y designado externamente. Esto tiene una doble connotación: por un lado, el objetivo del robot viene fijado por los humanos y, por otro, éste se vincula a una(s) finalidad(es) relativas al contexto práctico concreto donde ejerce las tareas. Así, deberíamos precisar que los robots sociales son *externalizaciones de una agencia interesada* y, por tanto, son agentes de naturaleza subrogada, que plantean interrogantes no solo sobre su ejecución de cursos de acción, sino prioritariamente sobre su razón de ser, es decir, de las finalidades e intereses de los que estos artefactos emergen, representan y a los que sirven.

De ese modo, pensar en términos de teleología e interés en materia de robótica social supone volver la atención ética hacia donde principalmente pertenece, esto es, la acción humana. Se trata de dos conceptos que deberían empezar a reivindicarse abiertamente en el marco de las reflexiones actuales sobre la actividad de este campo tecnocientífico.

Principalmente, se pueden alegar dos razones capitales a favor de la perspectiva del 'ser' como fundamento de una aproximación ética a la robótica social, que exponemos a continuación.

1) La coherencia con la naturaleza tecnológica de la robótica social

En primer lugar, poner el foco en la cuestión del 'ser' resulta coherente con la naturaleza tecnológica de la robótica social. Como expusimos antes, la finalidad última de esta actividad tecnocientífica siempre se vincula con los fines particulares del campo práctico donde se destinan los artefactos. En consecuencia, para la reflexión ética se hace necesario tener en cuenta las finalidades, valores y principios específicos a la esfera de actividad para la que los robots sociales se conciben como herramientas. Solo desde esta contextualización se puede deliberar críticamente sobre la pertinencia de sus funcionalidades, esto es, su quehacer.

Para esclarecerlo, tomemos la particularidad de los robots sociales, a saber, su capacidad para interactuar con las personas. Esta es su principal funcionalidad, objeto que requiere de examen ético. Ahora bien, dado que, por el carácter instrumental de la tecnología, la interacción como funcionalidad siempre se presentará como un medio para un determinado fin, una reflexión

ética consecuente deberá tomar en consideración el objetivo predeterminado de la interacción. Pues la problemática de la interacción humano-robot diferirá según su finalidad sea, por ejemplo, facilitar la ejercitación cognitiva del usuario en el ámbito terapéutico u ofrecer compañía en un contexto de asistencia sociosanitaria.

La perspectiva del ser conduce a ello de forma natural, pues parte de un interrogarse sobre el "sentido" del robot, atendiendo a la armonía entre sus funciones específicas y los objetivos de estas, dentro de un marco más amplio de finalidades, valores y hermenéuticas abiertas. Así, se garantiza una aproximación a la robótica social y sus artefactos centrada en su carácter *instrumental* respecto a algo concreto.

Esto significa atender al '*cómo*' que el artefacto representa con relación a unos objetivos. Con ello no nos referimos simplemente a avaluar si la robótica cumple *de facto*, a través de las funcionalidades desplegada en sus productos, las labores que pretendemos delegar. Es decir, no se trata de adoptar un paradigma restringido en que las finalidades de una actividad sean simplificadas a tareas. Para clarificarlo, aterramos un poco la cuestión: debería examinarse minuciosamente, por ejemplo, qué implica desarrollar artefactos robóticos para asistir en el comer. Dado que la finalidad es dar de comer, ¿pensaremos en términos de instrumentos que faciliten la ingesta, traduciendo el acto de ayudar a comer a garantizar el nutrirse? ¿Optaremos, entonces, por cubrir las necesidades de los pacientes que no pueden valerse por sí mismos a la hora de comer mediante el despliegue de un brazo robótico capaz de interacción física con el entorno, que le habilite para llevar a cabo su función sin peligro de dañar corporalmente a las personas? ¿O esto significaría dejar fuera la dimensión social del comer? ¿Desarrollamos mejor una aplicación robótica social, capaz de interacción "interpersonal", o entendemos que serán los humanos quienes complementarán la asistencia asegurando a la actividad el carácter relacional, que es interpersonal e impredecible? En caso que optemos por desarrollar artefactos sociales, ¿su interacción irá destinada, únicamente, a cumplir con la tarea de alimentar, o debería pensarse también desde la dimensión de los cuidados, del "estar al lado de", hacer compañía, y dar conversación o tacto? Tener en cuenta el estatuto de 'instrumento' de la robótica significa, fundamentalmente, tomar en consideración la forma en que sus artefactos ejecutan unas funciones –u, otramente dicho, la representación de valores y concepciones de la actividad que éstos personifican–. Pues debe haber consonancia entre el 'qué' se hace y el 'cómo' se lleva a cabo, en referencia a unos valores propios de la actividad.

En esta línea de pensamiento, ha surgido una interesante propuesta de aproximación normativa a la robótica para el ámbito de los cuidados, basada en el marco teórico de la "naturaleza de las actividades"(Santoni de Sio y van Wynsberghe, 2016). Se la conoce como "*nature-of-activities approach*"[31], dirigida, en este caso, a delimitar qué funciones delegar a los robots para los cuidados y cuáles no y cómo diseñarlos para que estas respondan a valores constitutivos del ejercicio de atención sanitaria. El mérito de la propuesta estriba en tomar en consideración la

[31] Podría discutirse la idoneidad del término "naturaleza" para denominar esta aproximación, por cuanto puede llevar a pensar que la propuesta de dichos autores se erige alrededor de una concepción de las actividades humanas como algo estático, inmodificable, con valores inmutables. Esto sería equivocado, pues de hecho se preocupan de explicitar una legítima diversidad interpretativa en torno a lo que significa exactamente una actividad, los valores asociados a ella. Por ello, sería una opción renombrar este enfoque como "aproximación a la dinámica de las actividades", haciendo hincapié en la idea de la actividad como un proceder lanzado a una constante regeneración y, por tanto, axiológicamente revisable.

distinción entre actividades "orientadas a la práctica" y "orientadas a objetivos" y partir de ella a la hora de reflexionar sobre qué porciones (tareas) constituyentes de la actividad del cuidado se podrían asignar al robot sin que, con ello, se dañe la dimensión que concerniría a los bienes internos de la misma (MacIntyre, 2019).

Resulta fundamental puntualizar que reflexionar éticamente desde una visión instrumental de la tecnología no implica desatender su carácter "mediador" (Verbeek, 2006). Es decir, examinar la coherencia de las funciones con las finalidades, en consonancia con valores definitorios de la actividad, no agota el objeto de reflexión ética. Pues los artefactos tecnológicos no son instrumentos neutrales, sino que condicionan decisivamente la forma en que se ejecutan las prácticas y, por tanto, su entramado de valores, por cuanto invitan a unas e inhiben otras posibilidades de "percepción y acción, experiencia y existencia" (Verbeek, 2006, 364). Este aspecto es crucial para la reflexión ética, porque anticipar qué tipo de mediación tecnológica queremos puede contribuir a desarrollar un tipo de robots con funcionalidades o características determinadas en lugar de otras, y todo ello puede aumentar o mermar la riqueza de los "entornos de funcionamiento" en que se introducen los artefactos (Toboso *et al.*, 2020).

2) Una mirada ética extensa más allá y acá de los impactos

En segundo lugar, adoptar la perspectiva del 'ser' posibilita una aproximación ética comprehensiva a la fuerza transformativa de esta actividad para la vida humana. Esto se explica por dos motivos.

Por un lado, porque al partir de unas coordenadas conceptuales que revelan a los robots como productos socioculturales –cuya razón de ser responde a un determinado sistema de valores y forma de entender y estructurar la vida en comunidad, esto es, a intereses humanos–, amplia la reflexión ética más allá de un mero ejercicio de "impactología". Es decir, evita reducirla exclusivamente a una avaluación crítica de los (predecibles) impactos de los robots.

En efecto, reducir a los impactos la avaluación ética de la tecnología entraña complejidades bien conocidas, como la imprevisibilidad de sus consecuencias (Jonas, 2015) o la cuestión sobre la perspectiva a partir de la cual éstas deberían estimarse (desde qué valores o principios, y en relación a quién, exactamente). Pero todavía más trascendente resulta el riesgo de caer en un "conservadurismo moral", al convertir la reflexión ética en una evaluación moral. Ciertamente, atender solo a los impactos producidos por la introducción robótica en nuestras prácticas podría precipitarnos a atrincherar nuestra aproximación en el marco de los valores actuales, sin atender la necesidad constante de repensarlos, tal como implica el ejercicio de la ética.

Bajo la perspectiva del 'ser', los robots sociales no son simplemente agentes artificiales que pueden asumir tareas esencialmente humanas, sino que representan una encarnación de cierta cosmovisión e intereses. Por ello, la reflexión no se limita a abordarlos como entidades "quienes" producen impactos en su quehacer, los cuales deben ser positivos –responsabilidad que, como dijimos, desde una posición ética desenfocada se suele interpretar como primariamente exigiendo desarrollar competencia moral en estos artefactos–. Contrariamente, se focaliza la atención antes que nada en la acción humana, tomando como objeto del análisis normativo el entramado de intereses, finalidades y valores a los que responden los productos de la actividad robótica. Esto supone, pues, no solo deliberar sobre 'medios' en un marco de 'fines' ya dados, sino abrir la consideración sobre 'qué medios' para '*qué fines*'.

Por otro lado, la perspectiva del 'ser' incluye en la reflexión ética una consideración a distintas esferas de la vida humana, más allá de la individual. Al centrar la atención primordialmente en los humanos –por ser los sujetos que deciden el 'para qué' del robot y que transforman el panorama de sus prácticas al incluir en ellas estas tecnologías–, la perspectiva del 'ser' no limita el ámbito de reflexión a la vida de los individuos concretos con quienes estos artefactos interactúan en el inmediato contexto de su uso, sino que también lo extiende a otras esferas, como la interpersonal, sectorial o institucional. De este modo, promueve una reflexión sobre el conjunto de nuestra forma de estructurar la vida humana, las prácticas y arquitecturas subyacentes a su lógica, esto es, una aproximación ética exhaustiva y, por tanto, madura. El interrogante está, en primer lugar, en qué tipo de prácticas (y valores) queremos fomentar y cuáles no –cuestión cuya importancia acertadamente enfatizan Sandel (2015) y MacIntyre (2001)–, y qué parte de nuestra agencia queremos delegar o no y porqué. Se trata, en definitiva, de un preguntarse por el tipo de sociedades y vida que queremos construir.

6. Conclusiones

El potencial robótico para asumir nuevos roles en prácticas humanas de dimensión relacional –es decir, en la esfera de las interacciones intersubjetivas– sitúa la robótica social como una de las grandes fuerzas disruptivas actuales de nuestra vida práctica. Esto urge a aproximarnos a dicha actividad tecnocientífica desde la disciplina de la ética. Sin embargo, para ello es importante establecer previamente los fundamentos de un pensamiento crítico-reflexivo bien orientado hacia este campo.

El presente artículo ha pretendido contribuir en esta línea, clarificando la perspectiva desde la que dirigir la mirada ética a la robótica social, que será aquella centrada primariamente en la cuestión del 'ser'. Así nos ubicamos dentro de las coordenadas conceptuales trazadas por las categorías de teleología e interés, que son clave para acercarnos correctamente al fenómeno nuclear de preocupación ética, relativo a la posible externalización de nuestra agencia en los robots sociales. Al volver el foco de atención ética a la acción humana, la 'perspectiva del ser' permite emprender una reflexión crítica verdaderamente capaz de orientar el potencial de (re)configuración de la vida humana que tiene la actividad tecnocientífica.

Agradecimientos

Este artículo ha sido posible gracias a la financiación recibida por parte del Ministerio de Ciencia e Innovación en el marco de las Ayudas para contratos predoctorales para la formación de doctores - FPI (PRE2018-084286), y por la Agencia Estatal de Investigación a través del Sello de Excelencia María de Maeztu del Institut de Robòtica i Informàtica Industrial (IRI), CSIC-UPC (MDM-2016-0656-18-2).

Reconocimiento

Quiero expresar mi gratitud a la Dra. Begoña Román, quien, con su generosidad intelectual, su entusiasmo profesional y su incansable predisposición al diálogo, ha contribuido inestimablemente a la publicación de este artículo.

Bibliografía

Allen, C., Wallach, W. and Smit, I. (2006) 'Why machine ethics?', *IEEE Intelligent Systems*, 21(4), pp. 12–17. doi: 10.1109/MIS.2006.83.

Andriella, A., Torras, C. and Alenyà, G. (2020) 'Cognitive System Framework for Brain-Training Exercise Based on Human-Robot Interaction', *Cognitive Computation*. doi: 10.1007/s12559-019-09696-2.

Baumer, E. P. S. and Silberman, M. S. (2011) 'When the implication is not to design (technology)', *Conference on Human Factors in Computing Systems - Proceedings*, pp. 2271–2274. doi: 10.1145/1978942.1979275.

Breazeal, C., Takanishi, A. and Kobayashi, T. (2008) 'Social Robots that Interact with People', in Siciliano, B. and Khatib, O. (eds) *Springer Handbook of Robotics*. Springer, pp. 1349–1369. doi: https://doi.org/10.1007/978-3-540-30301-5_59.

Brundage, M. (2014) 'Limitations and risks of machine ethics', *Journal of Experimental and Theoretical Artificial Intelligence*, 26(3), pp. 355–372. doi: 10.1080/0952813X.2014.895108.

Busquets Surribas, M. (2019) 'Descubriendo la importancia ética del cuidado', *Folia humanística*, (12).

Campa, R. and Campa, R. (2016) 'The rise of social robots : a review of the recent literature', *Journal of Evolution and Technology*, 26(1).

Camps, V. (2017) *Breve historia de la ética*. Barcelona: RBA Libros.

Cappuccio, M. L., Peeters, A. and McDonald, W. (2020) 'Sympathy for Dolores: Moral Consideration for Robots Based on Virtue and Recognition', *Philosophy & Technology*, 33(1), pp. 9–31. doi: 10.1007/s13347-019-0341-y.

Cave, S. *et al.* (2019) 'Motivations and Risks of Machine Ethics', *Proceedings of the IEEE*. Institute of Electrical and Electronics Engineers Inc., 107(3), pp. 562–574. doi: 10.1109/JPROC.2018.2865996.

Coeckelbergh, M. (2011) 'Is ethics of robotics about robots? Philosophy of robotics beyond realism and individualism', *Law, Innovation and Technology*, 3(2), pp. 241–250. doi: 10.5235/175799611798204950.

Coeckelbergh, M. (2012) 'Are emotional robots deceptive?', *IEEE Transactions on Affective Computing*. IEEE, 3(4), pp. 388–393. doi: 10.1109/T-AFFC.2011.29.

Coeckelbergh, M. *et al.* (eds) (2018) 'Envisioning Robots in Society - Power, Politics, and Public Space', in *Proceedings of Robophilosophy 2018 /TRANSOR 2018*. IOS Press. doi: 10.1017/CBO9781107415324.004.

Coeckelbergh, M. (2020) 'Should We Treat Teddy Bear 2.0 as a Kantian Dog? Four Arguments for the Indirect Moral Standing of Personal Social Robots , with Implications for Thinking About Animals and Humans', *Minds and Machines*. Springer Netherlands, (0123456789). doi: 10.1007/s11023-020-09554-3.

Cortina, A. (1996) 'El estatuto de la ética aplicada. Hermenéutica crítica de las actividades humanas', *Isegoría*, 13, pp. 119–134.

Cortina, A. (2007) *Ética mínima. Introducción a la filosofía práctica*. Tecnos. Madrid.

Echeverría, J. (2003) *La revolución tecnocientífica*. Madrid: Fondo de Cultura Económica de España.

Feil-Seifer, D. and Matarić, M. J. (2005) 'Defining Socially Assistive Robotics', in *9th International Conference on Rehabilitation Robotics*. IEEE, pp. 465–468.

Feil-Seifer, D. and Matarić, M. J. (2011) 'Socially Assistive Robotics: Ethical Issues Related to Technology', *IEEE Robotics and Automation Magazine*, 18(1), pp. 24–31. doi: 10.1109/MRA.2010.940150.

Fong, T., Nourbakhsh, I. and Dautenhahn, K. (2003) 'A survey of socially interactive robots', *Robotics and Autonomous Systems*, 42(3–4), pp. 143–166. doi: 10.1016/S0921-8890(02)00372-X.

Funk, M. and Coeckelbergh, M. (2020) *(Technical) Autonomy as Concept in Robot Ethics*, *Biosystems and Biorobotics*. Springer. doi: 10.1007/978-3-030-24074-5_12.

Gunkel, D. J. (2018) 'The other question: can and should robots have rights?', *Ethics and Information Technology*. Springer Netherlands, 20, pp. 87–99. doi: 10.1007/s10676-017-9442-4.

Johnson G., D. (2011) 'Computer Systems: Moral Entities but Not Moral Agents', in Anderson, M. and Anderson, S. L. (eds) *Machine Ethics*. Cambridge University Press, pp. 168–183. doi: https://doi.org/10.1017/CBO9780511978036.

Jonas, H. (2015) *El principio de responsabilidad. Ensayo de una ética para la civilización tecnológica.* Herder. Barcelona.

López Aranguren, J. L. (1994) *Ética. Obras completas, II.* Madrid: Trotta.

MacIntyre, A. (2001) *Animales racionales y dependientes. Por qué los seres humanos necesitamos las virtudes*. Barcelona: Paidós.

MacIntyre, A. (2019) *Tras la virtud*. Barcelona: Austral.

Millar, J. (2015) 'Technology as Moral Proxy: Autonomy and Paternalism by Design', *IEEE Technology and Society Magazine*, 34(2), pp. 47–55. doi: 10.1109/MTS.2015.2425612.

Noori, F. M., Uddin, Z. and Torresen, J. (2019) 'Robot-Care for the Older People: Ethically Justified or Not?', *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE, pp. 43–47. doi: 10.1109/DEVLRN.2019.8850706.

Nørskov, M., Seibt, J. and Santiago Quick, O. (eds) (2020) 'Culturally Sustainable Social Robotics', in *Proceedings of Robophilosophy 2020*. IOS Press.

O'Brolcháin, F. (2019) 'Robots and people with dementia: Unintended consequences and moral hazard', *Nursing Ethics*, 26(4), pp. 962–972. doi: 10.1177/0969733017742960.

Operto, F. and Veruggio, G. (2008) 'Roboethics: Social and Ethical Implications of Robotics', in

Siciliano, B. and Khatib, O. (eds) *Springer Handbook of Robotics*. Springer, pp. 1499–1524. doi: 10.1007/978-3-540-30301-5.

Payr, S. M. (2015) 'Towards Human-Robot Interaction Ethics', in Trappl, R. (ed.) *A Construction Manual for Robots' Ethical Systems. Cognitive Technologies*. Springer. doi: 10.1007/978-3-319-21548-8.

Ricoeur, P. (2008) *Lo justo 2. Estudios, lecturas y ejercicios de ética aplicada*. Madrid: Trotta.

Rodogno, R. (2016) 'Ethics and social robotics', *Ethics and Information Technology*. Springer Netherlands, 18(4), pp. 241–242. doi: 10.1007/s10676-016-9412-2.

Román Maestre, B. (2016) *Ética de los servicios sociales*. Herder.

Sandel, M. (2015) *Contra la perfección. La ética en la era de la ingeniería genética*. 2ª. Barcelona: Marbot Ediciones.

Santoni de Sio, F. and van Wynsberghe, A. (2016) 'When Should We Use Care Robots? The Nature-of-Activities Approach', *Science and Engineering Ethics*. Springer Netherlands, 22(6), pp. 1745–1760. doi: 10.1007/s11948-015-9715-4.

Sarrica, M., Brondi, S. and Fortunati, L. (2020) 'How many facets does a "social robot" have? A review of scientific and popular definitions online', *Information Technology and People*, 33(1), pp. 1–21. doi: 10.1108/ITP-04-2018-0203.

Seibt, J. (2017) 'Robophilosophy', in Braidotti, R. and Hlavajova, M. (eds) *Posthuman Glossary*. London: Bloomsbury Academic, pp. 390–393.

Seibt, J., Nørskov, M. and Schack Andersen, S. (eds) (2016) 'What Social Robots Can and Should Do', in *Proceedings of Robophilosophy 2016 / TRANSOR 2016*. IOS Press.

Sharkey, A. (2020) 'Can we program or train robots to be good?', *Ethics and Information Technology*. Springer, 22, pp. 283–295. doi: 10.1007/s10676-017-9425-5.

Sharkey, A. and Sharkey, N. (2012) 'Granny and the robots: Ethical issues in robot care for the elderly', *Ethics and Information Technology*, 14(1), pp. 27–40. doi: 10.1007/s10676-010-9234-6.

Sheridan, T. B. (2020) 'A review of recent research in social robotics', *Current Opinion in Psychology*. Elsevier Ltd, 36, pp. 7–12. doi: 10.1016/j.copsyc.2020.01.003.

Sparrow, R. and Sparrow, L. (2006) 'In the hands of machines? The future of aged care', *Minds and Machines*, 16(2), pp. 141–161. doi: 10.1007/s11023-006-9030-6.

Toboso, M. *et al.* (2020) 'Robotics as an Instrument for Social Mediation', *Biosystems and Biorobotics*, 25, pp. 51–58. doi: 10.1007/978-3-030-24074-5_11.

Tronto, J. (1993) *Moral Boundaries. A Political Argument for an Ethic of Care*. Routledge.

Vallès-Peris, N., Angulo, C. and Domènech, M. (2018) 'Children's Imaginaries of Human-Robot Interaction in Healthcare', *International Journal of Environmental Research and Public Health*. MDPI AG, 15(5). doi: 10.3390/ijerph15050970.

Vallor, S. (2015) 'Moral Deskilling and Upskilling in a New Machine Age: Reflections on the

Ambiguous Future of Character', *Philosophy and Technology*, 28(1), pp. 107–124. doi: 10.1007/s13347-014-0156-9.

Vandemeulebroucke, T., Dierckx de Casterlé, B. and Gastmans, C. (2018) 'The use of care robots in aged care: A systematic review of argument-based ethics literature', *Archives of Gerontology and Geriatrics*. Elsevier, 74(August 2017), pp. 15–25. doi: 10.1016/j.archger.2017.08.014.

Verbeek, P.-P. (2006) 'Materializing Morality. Design Ethics and Technological Mediation', *Science, Technology, & Human Values*, 31(3), pp. 361–380.

Veruggio, G. and Abney, K. (2014) 'Roboethics: The Applied Ethics for a New Science', in Lin, P., Abney, K., and A. Bekey, G. (eds) *Robot Ethics. The Ethical and Social Implications of Robotics*. Cambridge, Massachusetts: MIT Press.

Wallach, W. and Allen, C. (2009) *Moral Machines. Teaching Robots Right from Wrong*. Oxford University Press.

Weber, M. (2012) *El político y el científico*. Madrid: Alianza Editorial.

Žižek, S. (2013) *The Purpose of Philosophy is to Ask the Right Questions*, *Big Think*. Available at: https://bigthink.com/videos/the-purpose-of-philosophy-is-to-ask-the-right-questions.

2. The ethical issues of social assistive robotics: A critical literature review*

Júlia Pareto Boada
Institut de Robòtica i Informàtica Industrial, CSIC-UPC
Llorens i Artigas 4-6, 08028 Barcelona, Spain
jpareto@iri.upc.edu
ORCID iD: 0000-0003-4879-8800

Begoña Román Maestre
Facultat de Filosofia, Universitat de Barcelona
Montalegre 6, 08001 Barcelona, Spain
broman@ub.edu
ORCID iD: 0000-0001-6090-0172

Carme Torras
Institut de Robòtica i Informàtica Industrial, CSIC-UPC
Llorens i Artigas 4-6, 08028 Barcelona, Spain
torras@iri.upc.edu
ORCID iD: 0000-0002-2933-398X

Abstract: Along with its potential contributions to the practice of care, social assistive robotics raises significant ethical issues. The growing development of this technoscientific field of intelligent robotics has thus triggered a widespread proliferation of ethical attention towards its disruptive potential. However, the current landscape of ethical debate is fragmented and conceptually disordered, endangering ethics' practical strength for normatively addressing these challenges. This paper presents a critical literature review of the ethical issues of social assistive robotics, which provides a comprehensive and intelligible overview of the current ethical approach to this technoscientific field. On the one hand, ethical issues have been identified, quantitatively analyzed and categorized in three main thematic groups. Namely: Well-being, Care, and Justice. On the other hand –and on the basis of some significant disclosed tendencies of the current approach–, future lines of research and issues regarding the enrichment of the ethical gaze on social assistive robotics have been identified and outlined.

Keywords: Artificial intelligence, Care, Ethics, Healthcare, Human-robot interaction, Social assistive robotics, Justice, Well-being

1. Introduction

Plausibly, the branch of social robotics devoted to the development of assistive robots is the one that most clearly embodies the European ideal of an intelligent technology at the service of humans' well-being (High-Level Expert Group on AI, 2019). Indeed, engineering artificial intelligence (AI) tools for coping with the ontological condition of human vulnerability seems to be the highest exponent of a human-centric technology model aimed at prioritizing the empowerment of individuals for a higher quality of life.

This field is commonly known as "socially assistive robotics" (SAR). In general terms, it is focused on providing artificial intelligent robotic systems for aiding end-users with (physical or cognitive) special needs[32] in their daily activities (Tapus, Matarić and Scassellati, 2007). These vulnerable subjects include elderly adults, individuals with dementia, children with autistic spectrum disorders, convalescent patients, and people with other kinds of functional diversity needs. This

* **The text of this section entirely corresponds to the following publication: Pareto Boada, J., Román Maestre, B. and Torras, C. (2021) 'The ethical issues of social assistive robotics: A critical literature review',** ***Technology in Society, 67*****. doi: 10.1016/j.techsoc.2021.101726.**

[32] Although ultimately serving the needs of vulnerable end-users, SAR products are commonly conceived as tools for caregivers (primarily).

places SAR as a technoscientific field generally aimed at contributing to the practice of care (Feil-Seifer and Matarić, 2011). Specifically, socially assistive robots (SARs) are designed to support tasks in a broad range of care activities –like healthcare, physical and cognitive rehabilitation or therapy, domestic daily life and special education– and thus to be used in different settings –hospitals, elder-care facilities, homes and schools–. However, its defining particularity does not only lie in the type of tasks that they undertake (which are ultimately related to assistance[33]). It also consists in the way in which they carry them out: by means of socially interacting with humans –in virtue of which they can assume roles as coaching, motivating or providing company in ecosystems of care–. This is why SARs tend to be taxonomically understood as an intersection set between assistive robots, focused on assistive functions, and socially interactive robots, intended to interact with the human in a social way (Feil-Seifer and Matarić, 2005).

Despite the fact that "socially assistive robotics" has become the prevailing terminology, there are sound reasons to designate this field in broader terms as "social assistive robotics". First, the former label may not be inclusive enough (Payr, 2015): according to its original meaning (Feil-Seifer and Matarić, 2005), it leaves out the class of robots that, even if assisting through social interaction, may also involve some form of physical contact with users[34]. Second, the term "socially" is tautological, insofar as the so-called "socially assistive robots" are a subset of social robots. By definition, a social robot is an artificial intelligent entity (humanoid or other) that interacts in an "interpersonal manner" to achieve the predefined and domain-specific goals of the practical context in which it serves (Breazeal, Takanishi and Kobayashi, 2008). Thus, the concept of "social assistive robotics/robot" is comprehensive enough with the main acceptation that the term "socially" originally meant to stress about these intelligent systems: the act of assisting, primarily, through social (rather than physical) interaction. Therefore, we will employ the concept of "social assistive robotics" to refer to the branch of social robotics focused on assistance and "social assistive robots" to designate its products[35], alongside the widespread acronym SAR and SARs, respectively.

So far, the primary domains where SAR research is being applied belong to the healthcare field (Matarić, 2017). SAR mostly supports tasks aimed at helping vulnerable people in processes of restoration or health maintenance. Such tendency is in line with the European interest in healthcare as a major application area for robotics, AI, and digitalization development [8,9].

[33] For an insightful remark on the narrowed conception of assistance that is here at stake, see (Aparicio Payá *et al.*, 2019).

[34] This is the case, for instance, of the cognitive robotic system for assisting mild dementia patients in a brain-training exercise developed by (Andriella, Torras and Alenyà, 2020).

[35] A remark on the object of our review is here in order. Although some conversational AI-based systems which are already in the market, such as Amazon Alexa or Google Home, could develop some assistive tasks close to those of SARs, there are at least two significant differences between both kinds of artifacts that make it necessary to draw a distinction between them and grant a specific ethical attention to the latter, as done in this work. First, the embodiment of AI in a robotic entity entails an artificial agency whose ability to perform tasks within the physical world raises specific ethical challenges besides those posed by virtual conversational agents. Second, and more importantly, whereas the purpose of an AI conversational software product such as Amazon Alexa or Google Home can be defined by its user, SARs are conceived for and implemented within a practical context with its own (accepted and objective) ends. This requires of a very particular kind of ethical reflection, namely an exercise of applied ethics.

Although SARs' widespread implementation is (still) far from being a reality, significant European research initiatives (Fosch-Villaronga and Grau Ruiz. María Amparo, 2019) and pilot projects already being launched evince institutional prospects to incorporate these technologies within (healthcare) assistive contexts. An example of this is Barcelona's City Council pilot project to improve the quality of life of senior citizens through SARs (Ajuntament de Barcelona, 2020). This European goal can be explained as a response to the increasing populations with special needs (Matarić and Scassellati, 2016); a challenging phenomenon to a great extent due to the increase of ageing populations worldwide (United Nations, 2019), for which SARs are seen as a promising technological solution [9,12]. Indeed, aged care is becoming a SARs' central domain of application (Vandemeulebroucke, Casterle and Gastmans, 2020), which is why current ethical reflection on these intelligent robots focuses predominantly on their use in the context of elderly care.

In spite of its expectable significant contributions to care practice, SAR raises a considerable number of ethical challenges, given its disruptive potential for the organization (and conception) of our community life. Therefore, alongside its development, SAR is becoming a focus of growing ethical attention (Vandemeulebroucke, Casterle and Gastmans, 2020), already constituting a distinctive area of reflection within Roboethics (Tzafestas, 2018). Nevertheless, despite the widespread proliferation of ethical discussion on SAR, it is difficult to obtain a clear and comprehensive overview of the current debate. There is no way to straightforwardly get a complete outlook of the different ethical issues addressed in academic literature, nor to make outright sense of this fragmented ethical landscape –that is, to get an intelligible global picture of the concerns responding to some conceptual order.

There are several reasons behind the tangled ethical approach to SAR:

(1) Ethical reflection is dispersed throughout a heterogeneous body of literature. In part, this has to do with the inherent complexities of social robotics as an object of ethical reflection (Pareto Boada, 2021). For instance, the diversity of spheres of action involved in this technoscientific field implies the existence of multiple roots for SAR ethical issues, which may differ according to each developmental stage. Likewise, since the instrumental character of social robotics requires the contextualization of ethical reflection in the specific field of technological application, it fosters a diversification of the total body of SAR identifiable ethical issues into literature that concerns itself with different practical fields of assistance. In line with the current development trend, ethical reflection on SARs is mostly developed in healthcare robotics literature (Stahl and Coeckelbergh, 2016).

(2) There is a generalized terminological ambiguity regarding SARs, since they are very often designated with other concepts such as "care", "medical", or just "social robots" (Vallès-Peris and Domènech, 2020b, 2020a). Even if used interchangeably, these terms have relevant connotative differences, so this tendency hinders mapping the specific ethical issues of SAR.

(3) A general and thorough synthesis of SAR ethical issues so far addressed in literature is still lacking. Even though a valuable systematic review of ethics literature on SAR has been already conducted (Vandemeulebroucke, Dierckx de Casterlé and Gastmans, 2018), it is

narrow in terms of considered publications (fully elaborated argument-based literature), end-users (the elderly), and technological development stage (use).

This current scenario of the ethical debate on SAR is certainly problematic, since it may devalue the practical strength of ethics and its relevance for the normative guidance of technology's disruptive force already from the early and throughout all different stages of development[36]. To engage in a (much needed) fruitful and inclusive ethical dialogue (Fernández-Aller *et al.*, 2021) for the legitimate reconfiguration of human activity through SAR developments, it is necessary to start by putting in common, conceptually ordering and analyzing the ethical issues arguably associated with this branch of social robotics. Besides, having a whole understandable picture of the different ethical issues can foster the identification of potential future lines of research on SAR for human well-being.

To address this need, we conducted a literature review of SAR ethical issues, which had a twofold goal. On the one hand, to identify and analyze the different existing ethical concerns on SAR, in order to obtain an informed knowledge of the current landscape of scholar ethical reflection on this field of robotics. Thus, our first objective was to get a comprehensive view of the state of the art of ethical thinking on SAR, by seeking to answer two main questions: which are the ethical issues generally associated with SAR, and which of these are the most frequently addressed ones?

On the other hand, and on the basis of the detected tendencies in the academic ethical debate, the second goal was to identify and outline some lines of thought and issues that need to be developed further in future research to deepen and complete the ethical approach to SAR. Accordingly, we structured our research in two main stages: an extensive descriptive one –for which a quantitative and qualitative analysis was undertaken– and a germinal critical-reflective one devoted to assessing the literature review's results and sketching future directions for enriching ethical reflection on SAR.

2. Methodology

To identify and examine the main ethical issues associated with SAR, we conducted a literature review through the international bibliographic database of scientific journals Scopus[37] in July 2020. We focused exclusively on scholarly publications with the intent of offering an outlook on a well-supported ethical discussion in which concerns are grounded on up-to-date knowledge.

The bibliographical search was carried out through the following four terminological entries: (1) "ethics" AND "assistive robot*", (2) "ethics" AND "care robot*", (3) "ethics" AND "social robot*",

[36] The accelerated pace of technological development makes it highly important to foster not (only) a *post-facto* ethical reflection –thus focused on already designed products and its use–, but a proactive ethics instead (van Wynsberghe, 2016), engaged in all the different levels of the process and committed to key questions arising from the very same moment of conception, such as the teleological ones (why and what for). Whereas the former kind of ethical thinking would foster narrowing ethics to an exercise of impact assessment, the latter is crucial for the so-called Responsible Research and Innovation (Stahl and Coeckelbergh, 2016), and it is an indispensable element for the recently advocated "positive ethics" (Coeckelbergh, 2020) .

[37] Although the informational source of this literature review is limited to Scopus, this database is one of the major and most comprehensive ones (Pranckutė, 2021). This is why we consider our findings to be broad and representative enough for an overview of the current state of ethical scholarly thinking on SAR.

(4) "ethics" AND "human-robot interaction". The reason for this search was to broaden the scope of possible publications setting out the ethical issues of SAR, given the intersection of concepts that define the latter. First, "social assistive robot*" constitutes a searching subset of "assistive robot*", so the latter offers a richer literature niche regarding our aims. Second, given that "assistance" falls under the umbrella of "care practice" and that both terms are usually used in an overlapping sense in the academic literature on robotics, "care robot*" was a necessary search for the review's exhaustiveness. Third, as SAR constitutes a subset of social robotics, an entry regarding this broader field was appropriate. Lastly, SARs' distinctive feature is to support an assistive-related task through social interaction, which may have specific ethical implications that are not to be found in other assistive technologies.

The review process of this literature search (see Fig.1), data extraction, and quantification of ethical issues was completed through five steps.

First, for each one of the four searches' outcomes, we started by an initial screening of the publications' titles and abstracts in order to select and compile eligible literature for later full-text reading. Our preselection criterion was based on the fact that such sections had to indicate some consideration for the ethical implications of technology within the papers' content. To reach a broad overview on SAR ethical issues, the specific kind of SARs on which publications (might have) focused was not an exclusion criterion. Since social robotics is a quite recent technoscientific field of intelligent robotics (Mejia and Kajikawa, 2017), we did not apply any restrictive criterion regarding the publication period of potential papers, which oscillated, at the most, between 2004 and the first half of the year 2020. Only publications written in English were considered.

Secondly, we proceeded to a full-text reading of the preselected publications to carry out a double task: to refine the selection of relevant papers that fitted our research goal, and second, to inductively identify any ethical problems brought up by each of them and extract this information into a table.

Regarding the screening process by full-text reading, the applied inclusion criterion was that publications had to (1) address ethical issues[38] (2) related to social (assistive) robotics (3) with an *indirect* focus on healthcare. We interpreted the first requirement in a non-restrictive sense, meaning that the mere mention or reference to ethical issues was a criterion for inclusion. Thus, we did not limit the definitive selection of publications exclusively to papers that argumentatively engaged in ethical matters, exposing and arguing for a particular stance thereby. An enumeration or overview of ethical issues was also taken as a reason for inclusion. The second requirement responded to the fact that the object of ethical study is, taxonomically speaking, an intersection set. Thence, challenges may arise from the different aspects involved. The source of SAR ethical issues is a ramified one, rooted in the assistive dimension (and its

[38] The notions of ethical issues/challenges/concerns/problems/conflicts are indistinctively used in the current landscape of normative-oriented reflection on SAR. Here, we will embrace them all as synonyms under the broad concept of "ethical issues". However, we consider important to caution against the indiscriminate use of the notion "dilemma", which is also quite frequent in the ethics literature on SAR. Connotatively, this is a very limited word: it is too dualistic, since it captures an ethical issue in polarized terms. Not only is this an unusual form in which ethical challenges arise, but, most importantly, approaching reflection on SAR in these terms may impoverish the ethical scope of meaning of the issues at hand and the quality of our normative reflection regarding them.

conceptually related practice of care) and the robots' interactive functioning. We dismissed papers that explicitly excluded consideration to social robots (exclusion criterion 1), but not those focused on generic groups potentially involving them (as "assistive" or "care robots"). The reason for the third requirement is that we aimed at identifying a broad scope of ethical issues regarding SAR. Therefore, we did not want to narrowly restrict the search to a specific practical context of SARs' application (aged care, nursing care, mental or physical therapeutic care); nor to limit the end-users of the assistive practice to a particular profile of vulnerable individuals. However, since ethical reflection on technology requires a contextualized approach, we opted to select the major field of healthcare as the broad indirect focus of our search, that is, as the main field of assistance. By "indirect", we mean that even if we did not explicitly restrict the search to this field of application –thereby including papers that were unspecific about the practical context of assistance, and thus having a general focus on social settings–, we excluded publications explicitly centered on a field of application other than healthcare, like education (exclusion criterion 2). Our choice is consistent with the current research trends and expectations on SAR for healthcare, which makes it urgent to reflect on this matter ethically.

Papers dealing with research ethics were excluded (exclusion criterion 3), except those including reflection on other stages of activity regarding social (assistive) robotics. We limited the "snowball method" to an occasional use, specifically for papers where the ethical issues being introduced were directly enumerated from a secondary document included in their bibliography. For the data extraction in the table, duplicates were removed.

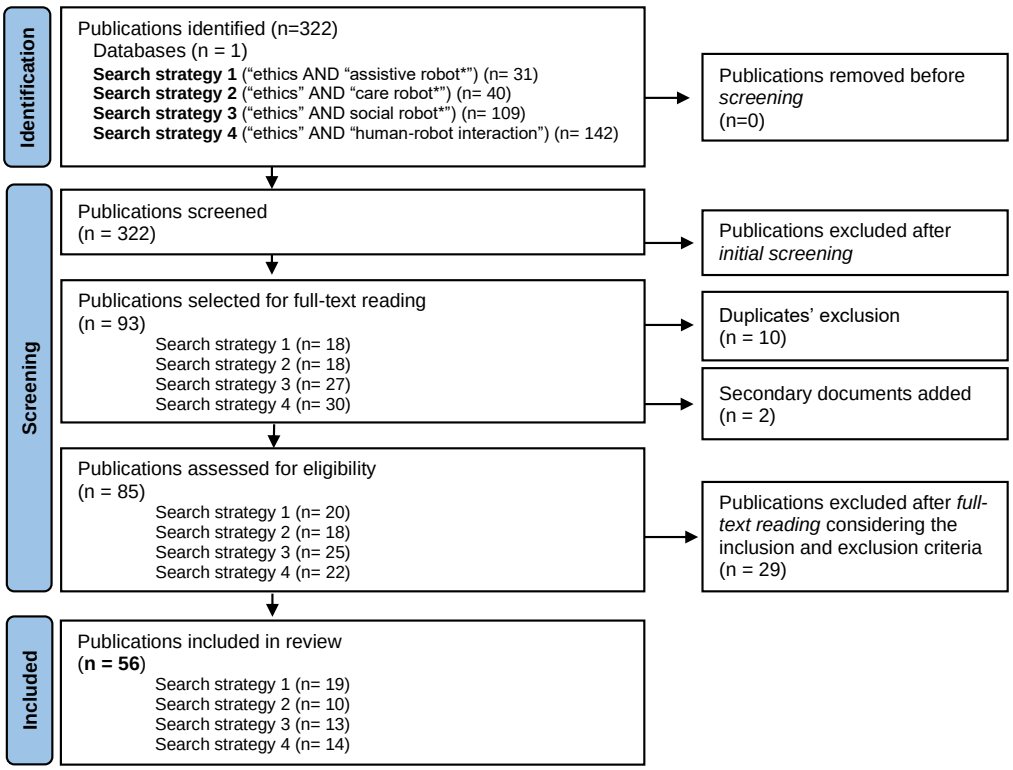

Fig. 1. Literature search and selection process. Source: Own elaboration based on Prisma 2020 (Page *et al.*, 2021).

Thirdly, from the table of ethical issues relative to each selected publication, we proceeded to make a comprehensive index of all the different ethical concerns raised in the reviewed

academic literature. Although most shared worries were exposed in the same or similar terms, we had to undertake a basic categorization to obtain a final list that was inclusive without overlapping implied connotations.

Fourthly, we compiled the indexed ethical issues on a spreadsheet alongside the corpus of selected publications (classified in four different subgroups according to their relative terminological search). This allowed us to use the index as an analytic tool for a second full-text reading of the papers. Through this, we re-identified the ethical issues they introduced, and marked them in the spreadsheet.

Lastly, each "SAR-associated" ethical issue was numerically quantified for the total corpus of the selected publications. To obtain an intelligible overall picture of the current landscape of ethical concerns on SAR, we categorized the identified issues in three thematic groups: Well-being, Care, and Justice. To that effect, we undertook a conceptual analysis on the content of the exposed issues.

3. Results

In our literature review, 56 publications were included for data extraction (Table 1), through which we identified a total of 26 ethical issues currently associated with SAR. These issues were very heterogeneous. Thus, for the quantitative analysis to be truly illustrative of the ethical debate on SAR, it had to be complemented with a categorization of the compiled concerns.

Table 1
List of included publications

Reference	Title
(Arnold and Scheutz, 2017)	Beyond Moral Dilemmas: Exploring the Ethical Landscape in HRI
(Battistuzzi *et al.*, 2018)	Embedding Ethics in the Design of Culturally Competent Socially Assistive Robots
(Battistuzzi *et al.*, 2020)	Socially Assistive Robots, Older Adults and Research Ethics: The Case for Case-Based Ethics Training
(Bisconti Lucidi and Nardi, 2018)	Companion Robots: The Hallucinatory Danger of Human-Robot Interactions
(Borenstein and Arkin, 2017)	Nudging for good: robots and the ethical appropriateness of nurturing empathy and charitable behavior
(Cappuccio, Peeters and McDonald, 2020)	Sympathy for Dolores: Moral Consideration for Robots Based on Virtue and Recognition
(Coeckelbergh, 2009)	Personal Robots, Appearance, and Human Good: A Methodological Reflection on Roboethics
(Coeckelbergh, 2011)	You, robot: on the linguistic construction of artificial others
(Coeckelbergh, 2015)	Artificial agents, good care, and modernity
(Coeckelbergh *et al.*, 2016)	A Survey of Expectations About the Role of Robots in Robot-Assisted Therapy for Children with ASD: Ethical Acceptability, Trust, Sociability, Appearance, and Attachment
(Damiano and Dumouchel, 2018)	Anthropomorphism in Human-Robot Co-evolution
(de Graaf, 2016)	An Ethical Evaluation of Human–Robot Relationships
(Dignum *et al.*, 2018)	Design for Values for Social Robot Architectures
(Feil-Seifer and Matarić, 2011)	Socially Assistive Robotics: Ethical Issues Related to Technology

(Fiske, Henningsen and Buyx, 2019)	Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy
(Haring *et al.*, 2019)	The Dark Side of Human-Robot Interaction: Ethical Considerations and Community Guidelines for the Field of HRI
(Heuer, Schiering and Gerndt, 2018)	Privacy and Socially Assistive Robots - A Meta Study
(Huber, Weiss and Rauhala, 2016)	The Ethical Risk of Attachment: How to Identify, Investigate and Predict Potential Ethical Risks in the Development of Social Companion Robots
(Ienca *et al.*, 2016)	Social and Assistive Robotics in Dementia Care: Ethical Recommendations for Research and Practice
(Jackson and Williams, 2019)	Language-Capable Robots may Inadvertently Weaken Human Moral Norms
(Koimizu, 2019)	Aged Care with Socially Assistive Robotics under Advance Care Planning
(Körtner, 2016)	Ethical challenges in the use of social service robots for elderly people
(Koyama, 2016)	Ethical Issues for Social Robots and the Trust-based Approach
(Lehoux and Grimard, 2018)	When robots care: Public deliberations on how technology and humans may support independent living for older adults
(Maalouf *et al.*, 2018)	Robotics in Nursing: A Scoping Review
(McBride, 2020)	Robot Enhanced Therapy for Autistic Children: An Ethical Analysis
(Miller, 2020)	Human Rights of Users of Humanlike Care Automata
(Misselhorn, Pompe and Stapleton, 2013)	Ethical Considerations Regarding the Use of Social Robots in the Fourth Age
(Noori, Uddin and Torresen, 2019)	Robot-Care for the Older People: Ethically Justified or Not?
(Nylander, Ljungblad and Jimenez Villareal, 2012)	A complementing approach for identifying ethical issues in care robotics – grounding ethics in practical use
(O'Brolcháin, 2019)	Robots and people with dementia: Unintended consequences and moral hazard
(Payr, 2015)	Towards Human-Robot Interaction Ethics
(Rabbitt, Kazdin and Scassellati, 2015)	Integrating socially assistive robotics into mental healthcare interventions: Applications and recommendations for expanded use
(Richardson, 2019)	The human relationship in the ethics of robotics: a call to Martin Buber's I and Thou
(Robson, 2018)	Intelligent machines, care work and the nature of practical reasoning
(Rodogno, 2016)	Social robots, fiction, and sentimentality
(Sharkey, 2014)	Robots and human dignity: a consideration of the effects of robot care on the dignity of older people
(Sharkey and Sharkey, 2012)	Granny and the robots: ethical issues in robot care for the elderly
(Sorell and Draper, 2017)	Second thoughts about privacy, safety and deception
(Sparrow, 2016)	Robots in aged care: a dystopian future?
(Sparrow, 2019)	Robotics Has a Race Problem
(Sparrow and Sparrow, 2006)	In the hands of machines? The future of aged care
(Stokes and Palmer, 2020)	Artificial Intelligence and Robotics in Nursing: Ethics of Caring as a Guide to Dividing Tasks Between AI and Humans
(Torras, 2019)	Social networks and robot companions: Technology, ethics, and science fiction
(Tzafestas, 2018)	Roboethics: Fundamental Concepts and Future Prospects
(Vallès-Peris, Angulo and Domènech, 2018)	Children's Imaginaries of Human-Robot Interaction in Healthcare

(Vallor, 2015)	Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character
(Vallverdú and Casacuberta, 2015)	Ethical and Technical Aspects of Emotions to Create Empathy in Medical Machines
(Van Aerschot and Parviainen, 2020)	Robots responding to care needs? A multitasking care robot pursued for 25 years, available products offer simple entertainment and instrumental assistance
(Van Maris *et al.*, 2020)	The Impact of Affective Verbal Expressions in Social Robots
(Vandemeulebroucke, Casterle and Gastmans, 2020)	Ethics of socially assistive robots in aged- care settings: a socio-historical contextualization
(Vandemeulebroucke, Dierckx de Casterlé and Gastmans, 2018)	The use of care robots in aged care: A systematic review of argument-based ethics literature
(Vanderelst and Winfield, 2018)	The Dark Side of Ethical Robots
(Weng and Hirata, 2018)	Ethically Aligned Design for Assistive Robotics
(Yew, 2020)	Trust in and Ethical Design of Carebots: The Case for Ethics of Care
(Zardiashvili and Fosch-Villaronga, 2020)	"Oh, Dignity too?" Said the Robot: Human Dignity as the Basis for the Governance of Robotics

Source: Own elaboration

The heterogeneity of ethical issues has to do with the variety of angles from which it can be critically reflected upon SAR. These giving rise to different types of concerns. Some of these angles configuring the ethical approach to SAR are the following: (1) ethical perspective (Vandemeulebroucke, Dierckx de Casterlé and Gastmans, 2018); (2) ontological assumptions –whether the focus is on the robot as an object or a "subject"–; (3) source of concerns –robot's particularities from which ethical problems arise, regarding both its technical elements (cameras, sensors, mobility, …) and functionalities or roles (specific tasks, social interactivity, autonomous decision-making, …)–; (4) contextualization of ethical reflection –whether reflection is led by ethical criteria belonging to a sole context (for instance: bioethical principles, happiness, or trust regarding the fields of healthcare, domestic life or institutions where technology is introduced), or to an intersection of contexts instead–; (5) stage of technological development –design, research, implementation or use.

We categorized the ethical issues in three thematic groups, which we have labeled as Well-being, Care, and Justice according to the ethical dimension of human life to which these are (allegedly) linked: individual, practice-related, and sociopolitical, respectively. That is, we classified them depending on the sphere of human life that is considered as the primary focus of SAR implications. For this classification, we remained faithful to the literature's underlying viewpoint, arranging the ethical issues in these groups according to how they are understood in the literature. This does not mean that these perspectives should not be further critically discussed: as set out in Section 4, it is important to broaden the meaning (and thus the ethical dimension) of some current concerns.

This categorization was deemed pertinent because it integrates the principal spheres of action coming into play with the introduction of social robots in assistive practical contexts, namely: (1) intersubjective/human-robot interaction; (2) (specific) human practice; (3) sociopolitical activity. The chosen (unrefined) terminology of Well-being, Care, and Justice aims at the inclusion of all these spheres of activity, thus respectively encompassing SAR ethical issues regarding its implications for (1) the individuals for which this technology is provided (users), (2)

the practice in which it is introduced and (3) society in its political structuring. These three categories are therefore related to three levels of ethical reflection.

We understand the notion of "practice" in MacIntyre's sense, namely, as "any coherent and complex form of socially established cooperative human activity through which goods internal to that form of activity are realized in the course of trying to achieve those standards of excellence which are appropriate to, and partially definitive of, that form of activity, with the result that human powers to achieve excellence, and human conceptions of the ends and goods involved, are systematically extended." (MacIntyre, 2007). Also, we regard care as the broad practice to which the activity of assistance contributes, and hence SAR. Therefore, with the use of the category of "Care" we mean to embrace ethical concerns on SAR regarding both the particularities of this relational human activity (goods, virtues, models of professional excellence…) and its (informal or institutional) organization –that is, the implications of SAR for the practical settings of assistance (distribution of tasks, institutional legitimacy and trust…).

The 26 identified ethical issues associated with SAR are shown in Fig. 2 alongside the results of their quantitative analysis (frequency); all of them correspondingly classified in the three main categories of Well-being, Care, and Justice. Notice that in a couple of cases (*) the same issue appears in different thematic groups (although with different frequency), given that it relates to more than one dimension of ethical concern.

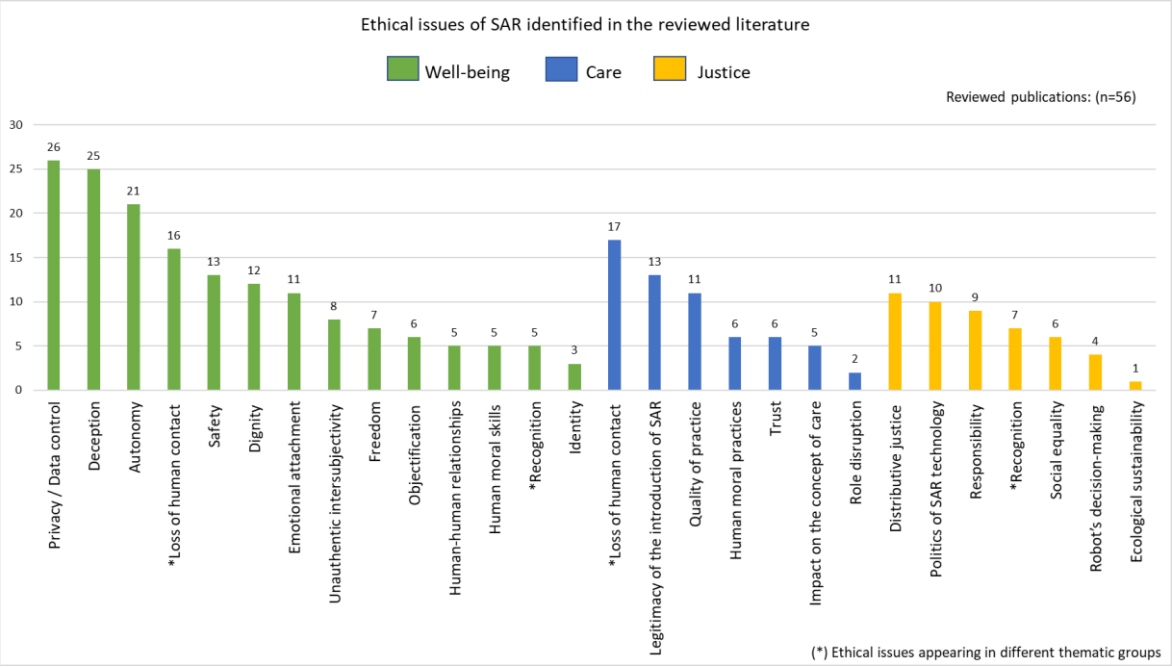

Fig. 2. Ethical issues associated with SAR in the reviewed literature. Source: Own elaboration.

As Fig. 2 reveals, among the 26 ethical issues of the current scholarly debate on SAR, the ones appearing most frequently are Privacy/Data Control (26 refs.), Deception (25 refs.) and Autonomy (21 refs.). In addition, since the mean number of appearances for each ethical issue is 10,42 times, these concerns constitute an outstanding focus of attention. In turn, all these three major concerns belong to the thematic group of Well-being, which, as Fig. 3 shows, gathers 60% of the identified ethical issues. It is followed by Care, which integrates 22% of them, and

Justice in the third place, with 18% over the total. Thus, the data shows a relevant tendency of ethical reflection on SAR, namely: that most ethical concerns have to do with SAR implications for the individual dimension of Well-being. Another significant fact highlighted in Fig. 2 is the disparity rate of ethical issues between Well-being and the other categories of Care and Justice.

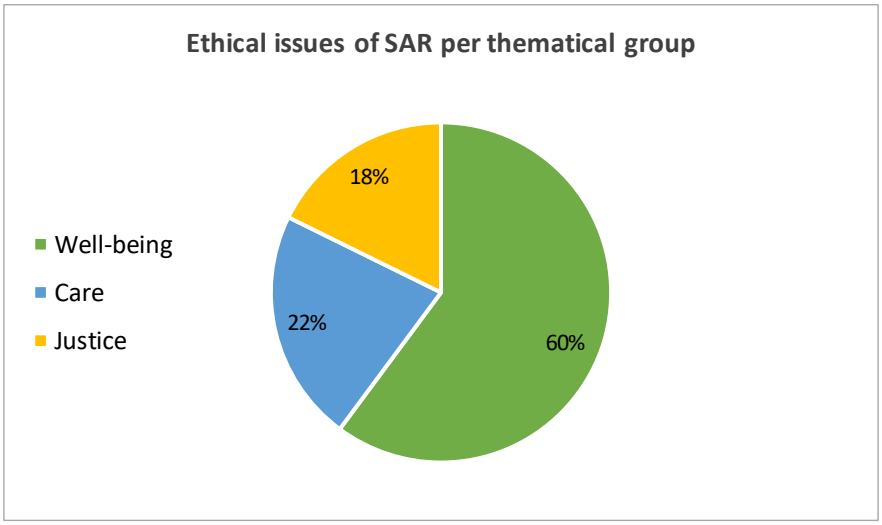

Fig. 3. Ethical issues of SAR per thematical groups. Source: Own elaboration.

3.1. Description of the identified ethical issues

Through our review process, we found out that, in many scholarly publications, ethical issues regarding SAR are barely developed. Ethical concerns are usually merely mentioned –without clarifying their meaning or the reasons for which they can be said to be posed by SAR– nor engaging or taking a stance in a further sustained argumentation on them. Besides, some of them are presented by using terms that stand as core ethical categories within our (Western) history of thought, the meaning of which is thus not only very rich but also disputed and permanently revisited. Therefore, in order to shed light on the current scholar ethical debate on SAR, a description of what is meant with each of the identified ethical issues is provided below according to what was concluded from our conceptual analysis of the literature[39].

3.1.1. Well-being

3.1.1.1. Privacy/Data Control. Privacy is the major issue of ethical concern in the literature on SAR. Although undefined in many publications, the concept seems to be generally understood according to Refs. (General Assembly of the United Nations, 1948) (The Member States, 2012): as a right against arbitrary interference with one's private life, implying a users' right to be in control of their personal information. Regarding this, a special matter of concern is the right to privacy of users with cognitive deterioration, which calls into question key related notions as "informed consent".

[39] In this descriptive section, no references regarding each of the ethical issues are systematically included. Whereas it could certainly provide a helpful relation of the topics and the corresponding literature, including references for each of the 26 points would entail lots of bibliographical repetitions, since most of the reviewed publications address a great number of the identified ethical issues. Inasmuch this would be quite displeasing for readers, it has been a discarded option.

SARs are considered a threat to privacy mainly in virtue of their monitoring capacity, through which they can collect, store, process and access personal data. Few authors delve further into the issue, by examining different types of privacy (Ienca *et al.*, 2016) and/or the link between different robotic technical features and the threat they pose to these (Heuer, Schiering and Gerndt, 2018). SARs' "intersubjective" capacity of interaction is also taken into account as a feature posing specific risks to privacy, being the creation of user profiles occasionally mentioned as a risk to privacy too. Sometimes, concerns on privacy relate to the feeling that users may have of being observed, of not being alone. Privacy can also be challenged by SAR's usual goal of watching over user safety, setting out a conflict of values requiring ethical ponderation.

3.1.1.2. Deception. Deception is a very contested issue in the literature on SAR. In broad terms, the concern lies in the deceitful intersubjective relationship that human-robot interaction (HRI) may entail. Beyond the dispute over whether social (assistive) robots are inherently deceptive or not, there is debate on whether deception is morally wrong or not. The former stance is advocated from two main perspectives.

On the one hand, deception is conceived as morally wrong in virtue of the negative consequences that it may have for users, especially vulnerable ones. For instance, deception could lead to over-trust the robot –thus endangering the user's safety or reasonable decision-making processes–, to foster the user's emotional dependency on the robot –by perceiving it as having feelings or emotional states–, or to rather trigger emotional discomfort –given the robot's incapacity for emotional reciprocity–. On the other hand, deception is argued to be intrinsically morally wrong, independently of whether upholding a false belief of the robot's (emotional) capacities or a distorted view of the interpersonal relationship with them has positive or negative consequences for the user. This stance is linked to a comprehension of delusion as inherently wrong, from which unauthentic relationships are deemed morally unacceptable (as the one held by an affective bond with an entity like a social robot).

Lastly, some authors link the moral wrongness of deception to the fact that, in itself, it is a product of human intent, blaming it on the deliberate goal of deceiving users that lies behind a robot's development –a deception that in many occasions is conceived as necessary in order to reach a robot's full functionality in an assistive context with vulnerable users–. In Kantian terms, this is a violation of human dignity because it involves instrumentalizing humans for the sake of achieving some goal.

3.1.1.3. Autonomy. (Human) autonomy is considered to be possibly challenged by SAR in numerous ways. From a consequential ethical perspective, an excessive (and thus inappropriate) extent or degree of assistance could cause a loss of the users' capacities alongside a dependency on technology; thereby conflicting with the primary goal of assistance (to promote autonomy) and conversely raising new forms of vulnerability. The problem lies in the correct suitability (in terms of the kind and the quantity or proportionality) of assistance. Other vulnerabilities of human autonomy may also be fostered by SARs, such as exposure to manipulation or improper decision-making delegation.

From another perspective, the infringement with autonomy is not a (potential) result but rather a constitutive fact of SARs. This view comes from a specific understanding of the politics of SARs

as a technology grounded in values and interests that are alien to that of ultimate users, which is why their introduction in the care practice equates to a violation of those users' capacity to live according to their own reasons. This relates to the concerns on *objectification* and *informed consent*.

Technological autonomy also triggers ethical concern on human autonomy, given the potential conflict that can take place between both, as in cases where a user's safety has to be balanced against their decision-making. This opens up the need of an ethical reflection upon the correct scope of beneficence (promotion of what is at the user's best interest).

3.1.1.4. Loss of human contact. SAR could foster or even increase social isolation, which is detrimental to one's well-being. First, SARs' introduction may come along with a reduction in human contact, thereby potentially lessening an end-user's opportunities for meaningful social interaction with humans. This is a problem both for individual patients and the organizational context in which the care practice is held. The shift in how the activity is carried out requires the adequate (re)articulating of the core values of the practice, all professionals involved, and that organization as a whole. Indeed, the problem has to do with the way in which robots are introduced: are they going to offer support within care practices as human replacements or as collaborative agents instead? And to assume which type of tasks and roles? Second, the worry about robots as an isolating factor is also linked to HRI, which can foster certain relationship pathologies, such as emotional dependence on the robot or the user's seclusion to their inner world.

3.1.1.5. Safety. SARs pose a potential threat to a human's physical and psychological integrity. This problem is usually introduced in terms of 'safety', which globally refers to the harmful effects that robots may have for users regarding this (double) aspect of well-being. However, current concerns on safety are still mostly explicitly directed at the potential harm to a user's physical integrity. That is, they relate to the risk of accidents arising from robots sharing the same space as humans and interacting with them. Regarding this, not only actual, but also 'perceived safety' is under consideration. In addition, the worry on safety is generally associated to the compliance with the (bioethical) principle of nonmaleficence, seemingly leaving out an explicit consideration to beneficence as part of that same concern (a robot's potential contribution to a good life). Safety is thus a concern highly related to responsibility as liability for damage.

Safety is not only under ethical consideration because of the (harmful) results of robotic functioning, but because of the possible tension that may arise between the user's safety (a major goal of certain applications of SARs) and their autonomy or privacy.

3.1.1.6. Dignity. Dignity plays a crucial role in the ethical debate on SAR: it is both a recurrent issue of direct concern –insofar as potentially compromised by SAR for different reasons–, and also the value on which other concerns are ultimately grounded –leading to either consequential or deontological assessments of SARs' introduction in social settings of human practices–. Therefore, there is a constellation of issues revolving around the core value of dignity, such as objectification, recognition, deception and identity, among others.

From a consequentialist point of view, SAR could both enhance or negatively impact human dignity. For instance, linking dignity to the capabilities approach, one perspective in this vein

argues to assess SAR depending on whether a robot's use expands or restricts the access to the basic set of capabilities to live a worthy human life. Also, dignity is under attack when these unauthentic intersubjective interactions alienate users from real life, thereby impoverishing their world. From a deontological perspective, SAR infringes upon human dignity because robots lack the basic abilities to provide assistance in an appropriate way, given that they cannot care and therefore cannot possibly fulfill the end-user's needs.

3.1.1.7. Emotional attachment. Basically, the reason why human emotional attachment to robots is introduced as an ethical issue is that it may generate contradictory effects with SARs' beneficial goal. For instance, it may lead to the loss of therapeutic benefits in the eventuality that the robot breaks down, does not fulfill the user's expectations or has to be shared with other users (feelings of jealousy). In the same lines, it may foster human over-trust on the robot with counterproductive results such as a robot's misuse by the user's over-delegation on it. Also, emotional attachment may promote certain pathologies of the human-robot relationship –such as a user's emotional dependency on the robot– and, in turn, human autonomy vulnerabilities.

Robot appearance is a matter of discussion closely linked to this issue, given that robotic design is considered to be a decisive element in the potential promotion of humans' emotional tie with robots.

3.1.1.8. Unauthentic intersubjectivity. The unauthentic intersubjectivity that characterizes HRI is a controversial subject of ethical debate, mostly regarded as morally problematic for several reasons. First, because it may reduce social life to an illusion, which comes along with the risk of an impoverishment of one's own world and life. The lack of a shared horizon of meaning between the human and the robot makes their relationship a mirroring one (of the human with themselves). Second, because the unauthentic intersubjectivity involved in HRI may affect the proper development or exercising of human moral faculties (empathy, care…), as well as foster an instrumentalist view of relationships, in which the "other" is not a "you" but rather an object. This endangers relations and forms of life that are intrinsically valuable and define us. Third, taken as a form of deception, unauthentic intersubjectivity is also argued to be wrong per se.

3.1.1.9. Freedom. The concern on freedom is grounded on a narrow understanding of this concept as "negative liberty" (Berlin, 1969), that is, as the absence of external restrictions to one's own activity. Under this view, SARs challenge a user's freedom insofar as their goal of promoting well-being (in terms of health) may require that these robotic devices conduct in a way that somehow restricts a human's actions or decisions. This concern is related to machine ethics, because it has to do with how the robot should ponder conflicting values in the course of selecting the conducts to undertake in relational practices.

3.1.1.10. Objectification. This issue mainly refers to two questions. On the one hand, to the feeling that SARs' end-users may experience when being assisted and interacting with robotic devices regarding vital human needs. On the other hand, it may also refer to the attitude or moral cosmovision underlying the decision of introducing social artificial agents to carry out care-related tasks –objectification thus describing the lack of respect to human dignity involved in such delegation–, as well as to the stigmatization of user collectives that robotic design and functionality may involve (for example, infantilization).

3.1.1.11. Human-human relationships. The effects that social robots may have on human-human relationships are an ethical concern linked to the one of human-moral skills. On the one hand, the kind of socialization that can be fostered by interacting with machines, and the changes in opportunities for people to practice certain social skills essential for human interaction, on the other, have deep implications for human-human relationships. These could be impoverished or limited by substituting humans for machines in the social interaction practice, which is argued to endanger both the intrinsic values of human relationships and our self-understanding as human beings.

3.1.1.12. Human moral skills. SAR is considered to basically pose a threat to human moral skills for two main reasons. The first has to do with HRI, whereas the second relates to the adoption of social robots in care. The reason for caring about moral skills is twofold: they are purported to be essential prerequisites for developing practical wisdom and virtuous character and, at the same time, to be intrinsically valuable. This last point is linked to the idea that moral deskilling implies diminished human beings (Vallor, 2015).

On the one hand, because of their "interpersonal" particular kind of interaction with humans, social robots may influence and shape human moral character, by potentially cultivating both vices and virtues. The source of ethical worry here is ultimately grounded on the relationship of pseudo-recognition involved in HRI. Human improper behavior and interaction with robots could foster a human's moral corruption. Even without explicit or intentional abusive behavior, the very same simulation of unconditional recognition carried out by robots could lead to a moral deskilling (for instance, it could normalize the experience of exercising control and power over what is seen as an autonomous agent with recognitive abilities).

On the other hand, the adoption of SARs in care, by outsourcing practices central to human existence to non-human actors, could blind us from the awareness of the constitutive vulnerability and (inter)dependance of human life, thus threatening the cultivation of virtues essential to a flourishing society. More specifically, the new technological practices could reduce the opportunities for cultivating moral skills regarding human caregiving.

A partly adjacent concern related to human moral deskilling is the influence that language-capable robots, because of their acting as social moral agents (and given their constrained dialogue systems), may have on human moral reasoning.

3.1.1.13. Recognition. The issue of recognition falls both within the individual and the sociopolitical dimension of ethical concern on SAR.

In general, the problem that SAR is considered to pose in terms of recognition has its source in a perspective focused on the HRI, in which the emphasized sphere of ethical attention is the individual one. The concern mostly arises from the characteristic ontology of the robot, which makes it unable to enter a genuinely affective relationship and, therefore, deprives the human interactant of recognition, which is a fundamental element of social relationships –which are in turn indispensable for well-being–. From a more social-relational point of view, it is argued that the unconditional recognitive relationship that the robot establishes with the human is a source of a human interactant's potential moral corruption, insofar as it makes the relational asymmetry be one of a power-relationship.

In fewer cases, SAR is understood to challenge recognition in that it may fail to respect the commitment with an individual's equal civic rights regarding politics of welfare. This would be the case if the needs, interests and (reasonable) preferences of assistive technology users were disregarded or not equally taken into account and represented by these technological developments. Whose priorities prevail, and how problems are defined, are matters of justice that have to do with the ethical issue of recognition[40]. Robot appearance may also have ethical implications for the sociopolitical dimension of recognition (see *Identity*).

3.1.1.14. Identity. As a matter of concern, identity is introduced in the literature on SAR in two senses. First, as a self-conception, thus having to do with the respect for oneself. SAR potentially challenges identity in terms of impinging on (a user's) comfort with one's own image. For instance, due to its technological design –which may reinforce the image of the specific impairment to which assistance is provided– or to the way in which the artifact assists the user –which can be considered to harm people's integrity–. Second, as an externally projected identity, thus relating to the image that third parties project onto users. In this sense, the concern has to do with the representation of an individual's or a collective's identity underlying SAR development, which can be an act of stigmatization not only impacting users individually, but also having ethical implications at a sociopolitical level. In turn, identity may be challenged by the inferences to which specific artifacts may lead to regarding aspects such as gender or race, which may be a collective's discriminatory representation infringing upon human dignity and equality. Reinforcing, at the same time, narrative and structural injustices. This concern is tightly linked to a robot's appearance.

3.1.2. Care

3.1.2.1. Legitimacy of the introduction of SAR. The legitimacy of the introduction and use of SAR in practical settings as a means of supporting their defining activity is a matter with key ethical significance at an organizational level, given that institutionalized practices must appropriately articulate the core values of professional, organizational and public ethics so as to grant a good service to citizenry. The question has to do mainly with two intertwined issues. On the one hand, with the consistency tool-task, i.e., whether (and how to ensure that) SARs are an appropriate tool for the task in which they are aimed to serve. This has to do both with the goals of SARs' function (whether these are reached or not) and the values that are essential to the SAR-assisted practice (i.e., the way in which it will be held by means of SARs' introduction). In turn, it implies an attention to the question of how to properly reshape the practical context. That is, how to redistribute the tasks or roles in order to guarantee an alignment with the specific practice's goals and values, as well as with the core civic ones (based on Human Rights). Closely tied to this first issue, SAR's legitimacy has also to do with the question whether a user's needs and preferences are actually fulfilled, since the consistency of tool-task partly depend on whether the tool is user-centered. In this sense, the so-called 'information gap' between technology design and the end-user's specific needs poses a great obstacle to legitimacy. Therefore, different stakeholders' involvement in technological development is often introduced as a linked ethical issue.

[40] Given the debate about the conceptual articulation of recognition and distribution in a theory of justice, we have decided to distinguish between both categories on this paper in order to leave it open for further exploration on how to connect them in regards to SAR ethical implications at a sociopolitical level.

Thus, the ethical issue of SAR's legitimacy has to do with an attention to goals, values and processes of technological development and implementation, and it connects to other ethical questions such as *responsibility* and *trust*.

3.1.2.2. Quality of practice. The implications that SARs entail regarding the quality of the practices that they are meant to support is an important focus of ethical concern. This worry is commonly expressed in terms of "dehumanization" of care practices and their settings, a phenomenon attributed to a robot's inability to enter in real intersubjective relationships with humans, which implies an inability to care (given their lack of moral agency and moral epistemology (Stokes and Palmer, 2020)). Ultimately, this concern is dependent on the redistribution of tasks; the central ethical question is how to reshape the traditional roles and functions of the professionals in the practice, i.e., which are the tasks that should be delegated to robots and why. Human substitution by robots could not only impact the practice quality, but also the meaning of care.

The quality of practice is also (partly) dependent on the consistency between the (presumed) tool and the task it is aimed to fulfill, which has to do both with the kind of task to be technologically assumed, and the way of carrying it out. To a large extent, such coherence depends on a proper knowledge and consideration to the context's particularities and the involved stakeholders. The endangering inconsistency tool-task could arise from a developer's knowledge gap about the needs and interests of the affected network of people, as well as the values and goals of the practice itself. Regarding this point, it is highly important to pay attention to the imaginaries of patients. This is mostly an unnoticed matter within ethical literature on SAR, the importance of which has been very well stressed out by Vallès-Peris, Angulo and Domènech (2018)[41].

3.1.2.3. Human moral practices. SAR may disrupt human moral practices that are constitutive of our societies and culture and, in turn, endanger both the internal goods of these practices, and certain human moral capacities that can only be developed and exercised through these forms of activity. SAR may erode care as a central practice of human moral life because it may reduce our engagement in such activity, thus lessening the cultivation of its associated moral skills and leading to a moral and professional deskilling. This has implications for the organizational sphere of human life, given that it challenges the core values of the exercised practice of care, which calls into question those of the whole institutional context within which is held.

3.1.2.4. Trust. SARs' introduction may distort the essential element of trust inherent to care relationships, which is a problem for the organizational context of the practice, since the quality of the institution (or service sector) requires ensuring the quality of the care practice. A robot's assumption of certain tasks that until now fell under (human) professionals' scope of action implies a restructuration of roles and responsibilities that may lead users to inappropriate levels of trust, both regarding SARs and human caregivers using them as a means of support within the practice (along with the organization as a whole). A major concern is to ensure that SARs are trustworthy, so that HRI's goals can be successfully achieved. Efforts are directed towards achieving social acceptance of robots. Besides the matter of safety and responsibility for harm,

[41] More recently, (Vallès-Peris and Domènech, 2020a) have further analyzed the role that roboticists' imaginaries play regarding this issue too.

the question of trust is also related to suitable knowledge about a robot's functions and capabilities (i.e., to a user's legitimate expectations), as well as to the coherence between SARs' functions and the practice's values and goals.

3.1.2.5. Impact on the concept of care. The new possibilities opened by SAR in the practice of care blur and call into question the meaning of such concept and its value and transcendence for human condition. In the literature, concerns on this issue relate to possible changes on social values surrounding care, as well as on society's concept of eldercare. Seemingly, the issue is linked to SARs particular ontology as 'almost-subjects', in virtue of which their inclusion in the practice of care disrupts our previous conceptualization of this activity as one exclusively entailing interpersonal relationships. Moreover, the concept of care may also change due to (new) needs either created or highlighted by these artifacts when introduced in the framework of social relations.

3.1.2.6. Role disruption. The introduction of SARs challenges current roles and responsibilities in care practice settings, thus threatening the quality of the practice and the essential element of trust that is constitutive of the relationship between caregivers and care recipients. Which tasks can be responsibly delegated to SARs or not in order to legitimately reshape these roles is a question with deep ethical implications at the organizational level.

3.1.3. Justice

3.1.3.1. Distributive justice. Some of the concerns on SAR revolve around distributive justice insofar as they have to do with the distribution of benefits and burdens across members of society. The matter is mainly about the fair allocation of SAR's initiatives' benefits and costs, being the latter primarily understood in terms of job impact (decrease of caregiver jobs due to the replacement of human workers by robots). The distribution of SARs and care as resources or goods is also an ethical issue falling under distributive justice concerns: who will have access to care robots? Could SAR contribute to a fairer distribution of care?

This issue is not argumentatively developed in the selected literature. The kind of individuals to whom these considerations of justice regarding SAR are meant to apply (whether among fellow citizens or rather international ones), is never explicitly stated, thus remaining unspecified whether distributive justice is contemplated within a certain political territory, or rather/also among countries, which would imply considerations of international distributive justice. Seemingly, it is the local distribution of benefits and costs that is under ethical reflection (with the exception of the question of ecological sustainability, which is presented as both a local and global matter of concern). Besides, distributive justice between generations is not mentioned in the literature, although SARs may well open up the need of reflecting upon the fair distribution of costs and benefits between contemporary and future generations (intergenerational justice).

3.1.3.2. Politics of SAR technology. This issue refers to the interests and values behind SAR initiatives and the question of their legitimacy. Concerns on SAR development as being driven by "technological solutionism" (Morozov, 2013) are quite frequent, and SARs' suitability as tools for solving social problems (as the shortage of available social services for the care of the elderly) is often called into question. Ethical reflection on this matter has to do with the need of examining and grounding the reasons for SAR initiatives. Which is the problem at which they aim to respond? Which are the underlying (economical, political, ideological) interests and how

are they being, or should be prioritized? Which are the values grounding our social practices around vulnerability, and how and to which extent are they re-configurated by SARs' introduction for care? The ethical issue thus revolves around the need of discussing, openly and inclusively at the societal level, the organization of the practice of care and the production of technological goods allocated to it. A usual worry regarding this issue is the prospect of the so-called "machinery of care".

3.1.3.3. Responsibility. SAR raises concerns on the ethical issue of responsibility, mainly in virtue of robots' technological autonomy, i.e., their ability to choose what to do based on previous information processing and regarding predefined goals, as well as their ability to behave accordingly. Therefore, the ethical worry mostly revolves around the question of liability for harm, where the latter is understood as the bad outcomes of SARs' functioning or tasks' execution. Who is ultimately responsible for the potential negative consequences of a robot's behavior, and who should be answerable to these? As a reflection linked to responsibility attribution, this issue is closely linked to matters of product safety and decision-making transparency of systems or "explainability" –which has to do with a key dimension of responsibility, namely: accountability (being able to explain and justify decisions).

3.1.3.4. Social equality. SAR has implications for social equality, since depending on how it is developed and implemented, it may either contribute to increase or lessen the equality of care both in terms of access and quality of treatment. Since intelligent autonomous machines are developed and trained using databases, social divide in terms of access is a big exclusion problem leading to inequality in the healthcare service (Vallverdú and Casacuberta, 2015), since the data of non-users won't be included into the databases from which the service is offered. This implies data bias by lack of representativity, which equates to an unequal (medical) treatment (or the impossibility of granting it) to such collectives.

3.1.3.5. Robots' decision-making. As a kind of technology to be introduced in daily life so as to autonomously carry out certain tasks within assistive practices, SARs' behavior has decisive consequences for individuals, which is normally why a robot's decision-making process is an issue of ethical concern. Guaranteeing that SARs will behave correctly according to the context goes beyond an issue of technical safety, and requires that their decision-making is aligned not only with the goals of their task, but also with human values. For it, a robot's ability to correctly assess and manage possible tensions or conflicts between different values that may arise in certain situations in real life (safety vs autonomy or privacy, for instance) seems to be a necessary (even though not a sufficient) condition. Therefore, this issue turns ethical attention to machine ethics, aimed at endowing robots with ethical reasoning capacities, so that their decision-making process is grounded on an understanding of, and an appropriate response to the moral relevant facts of each situation. Robot decision-making as a matter of ethical reflection on SAR is thus linked to concerns on *harm* and *responsibility*.

3.1.3.6. Ecological sustainability. The implications of SAR for both local and global sustainability are hardly ever examined in the literature. However, the supply of raw materials for robots, the energy consumption they require, and the dumping waste that these new care technologies generate are important ethical challenges of SAR. This issue is tightly related to matters of international distributive and intergenerational justice.

4. Reflections towards a critical approach to the ethical debate

Through the literature review, significant tendencies of the ethical approach to SAR have been disclosed, which evince a need of critically analyzing the way in which reflection is being directed. Accordingly, in this section we will outline some topics worth examining and discussing in view of enriching the ethical gaze on SAR.

(1) The individual-centered focus of ethical reflection

The ongoing ethical reflection on SAR predominantly focuses on the individual dimension of human life, i.e., on the implications that this technoscientific field has for individuals, who are, in turn, principally understood as SARs' users. Ethical approach to SAR thus takes what we have categorized as (individual) well-being as the primary dimension of normative concern. This is closely linked to the tendency to exclusively narrow the attention in the dyadic interaction between humans and robots, against which some authors have already argued (van Wynsberghe and Li, 2019) and which has been explicitly identified as a constitutive factor of a misguided ethical approach to social robotics (Pareto Boada, 2021).

The individual-centered perspective comes with a disproportionately fewer attention to SAR implications from the perspective of the (care) practice in which its artifacts are used, as well as from the macro sociopolitical level of justice. Descriptively, several reasons could be found behind this tendency[42]. However, from a normative-oriented point of view, what matters the most is that this is an important deficiency of the current ethical approach to SAR, which shows a continuation of the individualist assumptions and the "neglect of the political" underlying the mainstream philosophy of technology and ethics of technology (Coeckelbergh, 2018). We contend that this tendency should be overcome.

First and foremost, because an excessively restricted ethical focus at an individual level –at the expense of the two others– is symptomatic of a loss in perspective of the decisive interrelation between all three spheres of ethical import regarding SAR. It implies overlooking the conditioning that the sociopolitical structure has regarding the configuration of care practices and thus the influence that both of these have for an individual's life. Indeed, HRI in SARs' case

[42] First, such tendency reflects the kind of world where we find ourselves. A world in which technological development takes place within the frame of market dynamics, thereby narrowing ethical reflection on technology within its boundaries. This implies leaving aside an approach that thinks at a macro level, as an approach primarily led from a sociopolitical perspective would do. Second, ethical reflection on SAR must proceed as applied ethics that contextualizes reflection according to the values and goals of the specific practical field of activity for which technology is developed. This entails to circumscribe ethical reflection primarily within this specific (and already constituted) practice, which in a way means to endorse a conservative point of departure that may well lead to neglect matters related to the sociopolitical framework on which it takes place. The discipline of bioethics is an exemplary case of this: it also began its critical activity without calling into question the aforementioned macro level. Third, the fact that ethical reflection on SAR has been mostly led by a technoscientific professional profile could explain why the focus of ethical attention revolves around the user's (individual) well-being. Indeed, given the instrumental character of the technology, the prevailing ethics among engineers is a consequentialist one. That is, their ethical approach revolves around assessing the consequences of the artifacts in terms of meeting the expected goals in the established way. Insofar as SARs assist by interacting with humans, ethics is focused on the implications of this particular kind of robotic functioning. In addition, the fact that SARs are not still implemented at a large scale may well play a role in reinforcing this ethical perspective focused at the individual level, under which Care and Justice issues are much less addressed.

takes place within broader social practices that reflect values, goals and a specific cosmovision about how to organize human life. From a holistic consideration on the person, individuals must be taken in their situatedness. Hence, the constitutive interrelation between (individual) Well-being and the spheres of Care and Justice cannot justifiably be disregarded. This would mean to neglect the role of sociopolitical structure in easing or hindering, to a greater or lesser extent, the possibility of covering needs and developing personal autonomy.

Moreover, insofar as ethical reflection on SAR must primarily proceed as applied ethics understood as critical hermeneutics of human activity (Cortina, 1996), if anything, it is the (care) practice sphere the capital one. SAR should be primarily approached in light of the specific practice at which it aims to serve, within which individuals are not mere monads but members of a relational network of human activity.

Finally, an ethical approach that leaves insufficiently unattended SAR disruption potential regarding other dimensions of human life besides the (interactants') individual one, may come along with the risk of converting ethical reflection into a mere exercise of moral evaluation within an unrevised framework of values and (given) ends, by overlooking the question about the type of practices and societies we actually want, and how to accordingly (re)configurate life through SAR.

(2) Teleological and anthropological assumptions of SAR

The current ethical approach to SAR generally lacks reflection and discussion on SAR teleology, that is, on conceptual assumptions on "assistance", "care" and other correlated notions (human well-being, human capabilities, autonomy…) that underlie this field's development. The constellation of SAR teleological-related meanings should be examined, since it is always linked to a particular anthropology that should be carefully analyzed and further discussed. For now, SAR development hints at the specific idea of human vulnerability and fragility as an annoyance, the care of which can be delegated to technology. The background anthropology is thus a liberal one, revolving around capacitism and adultism, from which SAR is ultimately aimed at replacing capacities.

(3) Restricted understanding of core ethical concepts

The literature review reveals a very narrowed understanding of some ethical concepts around which SAR's problems stand, which impoverishes the ethical approach to the disruptive implications of this technoscientific field. Indeed, the limitation in such notions' scope of meaning correlates to a loss of sight of the interconnection between the three main spheres of ethical concern (individual, care practice-related, and sociopolitical).

For instance, (human) freedom is currently understood as what is philosophically known as "negative liberty" (Berlin, 1969). The ethical implications of SAR regarding this issue, though, could and should be broadened up by approaching the matter from a deeper understanding of this notion in its dimension of "positive liberty" (Berlin, 1969). That is, closer to freedom as autonomy –in which freedom has to do with self-realization, with taking over the own life–. It could also be interesting to think from the perspective of Pettit's "republican conception of freedom as non-domination" (Pettit, 2002). Delving into the meaning of this notion would definitely allow for a richer normative-oriented reflection on SAR that takes into account the

interdependence existing between freedom and the sociopolitical structuring of human life, thus approaching SAR's power of domination both at an interpersonal and structural level.

The same happens with responsibility, which is generally understood as liability for harm and thus revolves around the distribution of duties to answer for bad outcomes, which, on top of that, are linked to the (AI-based) robot's behavior. A more "substantial" concept of responsibility (Jonas, 2015) would enrich ethical reflection, by enabling us to think in terms of accountability for the development of SAR (teleology, interest), which implies taking a sociopolitical perspective that approaches the matter in light of justice.

Also privacy is misguidedly understood in the current debate on SAR in too individualistic terms, although the collective ethical dimension of privacy has been already well highlighted (Véliz, 2020). The implications of SAR for privacy are mainly thought within the frame of a robot's impact on their interactant's life. It should be considered to reframe reflection on SAR's threat to privacy from a sociopolitical point of view, that is, in terms of justice.

(4) Overlooked ethical issues in Well-being

Within the constellation of concerns related to (individual) Well-being, some important issues are overlooked. Since current ethical reflection on SAR is primarily focused on this sphere (due to HRI's central place within the ethically scrutinized spheres of activity impacted by SAR), engaging in those missing issues could significantly contribute to enrich the state of the art.

(i) Distinction between privacy, intimacy and interiority. Although Privacy is the most commonly addressed ethical issue on SAR, there is no attention to other notions that are related to its semantic field (Schoeman, 1984; Torralba, 2009) and incorporate nuances worth examining (Illa Mestre, 2018) regarding the implications of SAR at an individual level of the user's life.

(ii) The individual's "possibility to be". At the level of HRI, there are serious ethical issues (besides deception) alarmingly missing in the current landscape of reflection. These are ultimately related to the implications of SAR for the individual's "possibility to be" (and not only "to do"), in virtue of the standardization of people and relations that HRI entails.

In relational practices of care, HRI may condemn the "being" by reducing human interactants to a specific existential dimension, according to the user model that the individual represents. The interactant "stops" being a person and becomes a "model" within the assistive relationship. In turn, this model further reduces the person to quantifiable and operable patrons –sentencing them to remain as according to what they are said to be–. This latter point is related to the paradigm of experience on which a robot's learning and interaction takes place, which has, until now, been disregarded in relation to SAR –although the ethical implications of the algorithmic functioning and decision-making are indeed taken into account in other application domains of AI-systems.

Some features of the current debate landscape on SAR explain the thematic oversight on the individual's "possibility to be"; such as the above highlighted restricted view on the key concepts of autonomy and freedom, as well as the scarce and poor reflection on the notions of identity

and recognition, and the absence of an examination of concepts like experience, domination and difference. Therefore, a philosophical anthropology of (inter)subjectivity would be in order, since it would advance some nuclear concepts regarding the human condition (unpredictability, affectability, openness to future) that are essential to reveal and comprehensibly address the ethical challenges of HRI.

5. Conclusions

This literature review has provided an overall and hopefully clarifying picture of the current ethical reflection on SAR, thus tackling the until now little structured landscape of ethical issues associated to this technoscientific field of intelligent robotics. Thereby, some relevant tendencies and thematic deficits of the current approach have been identified and outlined, which should be further critically examined and addressed. Hence, this review opens up and points at new research directions on SAR ethical implications. It constitutes a basis on which to build a research agenda for widening and deepening the normative thinking on this technoscientific field of activity. To contribute to such agenda, we sketch below some of the future lines of research that may arise from our critical literature review.

- To extend the conceptual framework from which to comprehensively identify and analyze SAR ethical implications.

- To delve into particular concepts belonging to philosophical disciplines (e.g., freedom or responsibility) and integrate other notions that may also be relevant for addressing the ethical implications of SAR (e.g., experience, intimacy).

- To consider the identified ethical issues of SAR by drawing upon conceptual resources of political philosophy and philosophical anthropology.

- To examine SAR teleological and anthropological assumptions and relate them to the different existing theoretical frameworks on care. This theoretical work could latter provide practical coordinates for grounding decisions on the different stages of SAR development. For instance, regarding the question on which services and roles should be redistributed through SARs' deployment, under which conditions and how; as well as on the elements that engineers should take into account in order to serve the aims and values of the practice to which they give support to through SAR solutions.

- To integrate a critical theory perspective in the ethical approach to SAR, in order to include a teleological reflective momentum that discusses the ends, beyond the means, and to engage in the question about the type of practices, societies and lives that we want to (re)configurate through the design, deployment and usage of SAR solutions.

- To delve into other ethical challenges of HRI besides deception, with an special focus on the implications that the algorithmic functioning of SARs may have for the well-being of its interactants.

- To set practical principles for SAR deployment (design, research, implementation and use) under the light of the three interrelated categories of Well-being, Care and Justice. To that end, and given that SARs are aimed at supporting practices of care, existing frameworks of applied ethics (such as the principles of bioethics) should be revisited and updated considering the novelty of SARs-assisted care practice.

- To examine the relevance of cultural aspects regarding SAR development by drawing upon the difference between the ethics of minima and maxima.

In conclusion, this literature review is a first stage of a larger research process aimed at contributing to the ethical debate on SAR by completing the scope of issues that should be taken into account, and delving further into the most normatively relevant ones.

Funding

This work has been partially supported by the Spanish Ministry of Science and Innovation under a FPI scholarship for predoctoral contracts for the training of doctors (PRE2018-084286), and by the Spanish State Research Agency through the María de Maeztu Seal of Excellence to IRI (MDM-2016-0656).

Author's contributions/statement

All authors contributed to this study. The literature search and data analysis were performed by Júlia Pareto Boada. The first draft of the manuscript was written by Júlia Pareto Boada and all authors commented on and critically revised the previous versions of the manuscript. All authors read and approved the final manuscript.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Appendix A. Supplementary data

Suplementary data to this article can be found online at https://doi.org/10.1016/j.techsoc.2021.101726

References

Van Aerschot, L. and Parviainen, J. (2020) 'Robots responding to care needs? A multitasking care robot pursued for 25 years, available products offer simple entertainment and instrumental assistance', *Ethics and Information Technology*. Springer Netherlands, (0123456789). doi: 10.1007/s10676-020-09536-0.

Ajuntament de Barcelona (2020) *Misty II the social robot becomes part of the lives of twenty senior citizens*. Available at: https://www.barcelona.cat/infobarcelona/en/tema/senior-citizens/misty-ii-the-social-robot-becomes-part-of-the-lives-of-twenty-senior-citizens_907645.html (Accessed: 31 July 2021).

Andriella, A., Torras, C. and Alenyà, G. (2020) 'Cognitive System Framework for Brain-Training Exercise Based on Human-Robot Interaction', *Cognitive Computation*. doi: 10.1007/s12559-019-09696-2.

Aparicio Payá, M. *et al.* (2019) 'Un marco ético-político para la robótica asistencial. An Ethical-Political Framework for Assistive Robotics', *ArtefaCTos. Revista de estudios de la ciencia y la tecnología*, 8(1), pp. 97–117.

Arnold, T. and Scheutz, M. (2017) 'Beyond Moral Dilemmas: Exploring the Ethical Landscape in HRI', *ACM/IEEE International Conference on Human-Robot Interaction*, Part F1271, pp. 445–452. doi: 10.1145/2909824.3020255.

Battistuzzi, L. *et al.* (2018) 'Embedding Ethics in the Design of Culturally Competent Socially Assistive Robots', *IEEE International Conference on Intelligent Robots and Systems*, pp. 1996–2001. doi: 10.1109/IROS.2018.8594361.

Battistuzzi, L. *et al.* (2020) 'Socially Assistive Robots, Older Adults and Research Ethics: The Case for Case-Based Ethics Training', *International Journal of Social Robotics*. Springer Netherlands. doi: 10.1007/s12369-020-00652-x.

Berlin, I. (1969) 'Two Concepts of Liberty', in *Four Essays on Liberty*. Oxford University Press.

Bisconti Lucidi, P. and Nardi, D. (2018) 'Companion Robots: The Hallucinatory Danger of Human-Robot Interactions', *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 17–22. doi: 10.1145/3278721.3278741.

Borenstein, J. and Arkin, R. C. (2017) 'Nudging for good: robots and the ethical appropriateness of nurturing empathy and charitable behavior', *AI and Society*. Springer London, 32(4), pp. 499–507. doi: 10.1007/s00146-016-0684-1.

Breazeal, C., Takanishi, A. and Kobayashi, T. (2008) 'Social Robots that Interact with People', in Siciliano, B. and Khatib, O. (eds) *Springer Handbook of Robotics*. Springer, pp. 1349–1369. doi: https://doi.org/10.1007/978-3-540-30301-5_59.

Cappuccio, M. L., Peeters, A. and McDonald, W. (2020) 'Sympathy for Dolores: Moral Consideration for Robots Based on Virtue and Recognition', *Philosophy and Technology*. Philosophy & Technology, 33(1), pp. 9–31. doi: 10.1007/s13347-019-0341-y.

Coeckelbergh, M. (2009) 'Personal Robots, Appearance, and Human Good: A Methodological Reflection on Roboethics', *International Journal of Social Robotics*, 1(3), pp. 217–221. doi: 10.1007/s12369-009-0026-2.

Coeckelbergh, M. (2011) 'You, robot: On the linguistic construction of artificial others', *AI and Society*, 26(1), pp. 61–69. doi: 10.1007/s00146-010-0289-z.

Coeckelbergh, M. (2015) 'Artificial agents, good care, and modernity', *Theoretical Medicine and Bioethics*. Kluwer Academic Publishers, 36(4), pp. 265–277. doi: 10.1007/s11017-015-9331-y.

Coeckelbergh, M. *et al.* (2016) 'A Survey of Expectations About the Role of Robots in Robot-Assisted Therapy for Children with ASD: Ethical Acceptability, Trust, Sociability, Appearance, and Attachment', *Science and Engineering Ethics*. Springer Netherlands, 22(1), pp. 47–65. doi:

10.1007/s11948-015-9649-x.

Coeckelbergh, M. (2018) 'Technology and the good society: A polemical essay on social ontology, political principles, and responsibility for technology', *Technology in Society*. Elsevier Ltd, 52, pp. 4–9. doi: 10.1016/j.techsoc.2016.12.002.

Coeckelbergh, M. (2020) *AI Ethics*. MIT Press.

Cortina, A. (1996) 'El estatuto de la ética aplicada. Hermenéutica crítica de las actividades humanas', *Isegoría*, 13, pp. 119–134.

Damiano, L. and Dumouchel, P. (2018) 'Anthropomorphism in Human-Robot Co-evolution', *Frontiers in Psychology*, 9(MAR), pp. 1–9. doi: 10.3389/fpsyg.2018.00468.

Dignum, V. *et al.* (2018) 'Design for Values for Social Robot Architectures', *Frontiers in Artificial Intelligence and Applications*, 311(January 2019), pp. 43–52. doi: 10.3233/978-1-61499-931-7-43.

Dolic, Z., Castro, R. and Moarcas, R. (2019) *Robots in healthcare: a solution or a problem?, Study for the Committee on Environment, Public Health, and Food Safety*, *European Parliament*.

European Commission (2020) *White Paper on Artificial Intelligence - A European approach to excellence and trust*. doi: 10.1017/CBO9781107415324.004.

European Parliament (2017) 'Civil Law Rules on Robotics European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)', (July 1985), p. 23. Available at: http://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.pdf.

Feil-Seifer, D. and Matarić, M. J. (2005) 'Defining Socially Assistive Robotics', in *9th International Conference on Rehabilitation Robotics*. IEEE, pp. 465–468.

Feil-Seifer, D. and Matarić, M. J. (2011) 'Socially Assistive Robotics: Ethical Issues Related to Technology', *IEEE Robotics and Automation Magazine*, 18(1), pp. 24–31. doi: 10.1109/MRA.2010.940150.

Fernández-Aller, C. *et al.* (2021) 'An Inclusive and Sustainable Artificial Intelligence Strategy for Europe Based on Human Rights', *IEEE Technology and Society Magazine*, (March). doi: 10.1109/MTS.2021.3056283.

Fiske, A., Henningsen, P. and Buyx, A. (2019) 'Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy', *Journal of Medical Internet Research*, 21(5), pp. 1–12. doi: 10.2196/13216.

Fosch-Villaronga, E. and Grau Ruiz. María Amparo (2019) 'Expert Considerations for the Regulation of Assistive Robotics. A European Robotics Forum Echo', *Dilemata, Revista Internacional de Éticas Aplicadas*, (30), pp. 149–169.

General Assembly of the United Nations (1948) *Universal Declaration of Human Rights*.

de Graaf, M. M. A. (2016) 'An Ethical Evaluation of Human–Robot Relationships', *International Journal of Social Robotics*. Springer Netherlands, 8(4), pp. 589–598. doi: 10.1007/s12369-016-

0368-5.

Haring, K. S. *et al.* (2019) 'The Dark Side of Human-Robot Interaction: Ethical Considerations and Community Guidelines for the Field of HRI', *14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, pp. 689–690. Available at: http://arxiv.org/abs/1504.04339.

Heuer, T., Schiering, I. and Gerndt, R. (2018) 'Privacy and Socially Assistive Robots - A Meta Study', in *Privacy and Identity Management. The Smart Revolution*. Springer International Publishing, pp. 265–281. doi: 10.1007/978-3-319-92925-5.

High-Level Expert Group on AI (2019) 'Ethics Guidelines for Trustworthy AI'. European Commission, pp. 1–41.

Huber, A., Weiss, A. and Rauhala, M. (2016) 'The Ethical Risk of Attachment: How to Identify, Investigate and Predict Potential Ethical Risks in the Development of Social Companion Robots', *ACM/IEEE International Conference on Human-Robot Interaction*, 2016-April, pp. 367–374. doi: 10.1109/HRI.2016.7451774.

Ienca, M. *et al.* (2016) 'Social and Assistive Robotics in Dementia Care: Ethical Recommendations for Research and Practice', *International Journal of Social Robotics*. Springer Netherlands, 8(4), pp. 565–573. doi: 10.1007/s12369-016-0366-7.

Illa Mestre, M. (2018) *Proposta d'una polisèmia estructurada del concepte <<Intimitat>>*. Universitat de Barcelona.

Jackson, R. B. and Williams, T. (2019) 'Language-Capable Robots may Inadvertently Weaken Human Moral Norms', *ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 2019-March, pp. 401–410. doi: 10.1109/HRI.2019.8673123.

Jonas, H. (2015) *El principio de responsabilidad. Ensayo de una ética para la civilización tecnológica.* Herder. Barcelona.

Koimizu, J. (2019) 'Aged Care with Socially Assistive Robotics under Advance Care Planning', *Proceedings of IEEE Workshop on Advanced Robotics and its Social Impacts, ARSO*. IEEE, 2019-Octob, pp. 34–38. doi: 10.1109/ARSO46408.2019.8948742.

Körtner, T. (2016) 'Ethical challenges in the use of social service robots for elderly people', *Zeitschrift fur Gerontologie und Geriatrie*, 49(4), pp. 303–307. doi: 10.1007/s00391-016-1066-5.

Koyama, T. (2016) 'Ethical Issues for Social Robots and the Trust-based Approach', *Proceedings of IEEE Workshop on Advanced Robotics and its Social Impacts, ARSO*. IEEE, 2016-Novem, pp. 1–5. doi: 10.1109/ARSO.2016.7736246.

Lehoux, P. and Grimard, D. (2018) 'When robots care: Public deliberations on how technology and humans may support independent living for older adults', *Social Science and Medicine*. Elsevier, 211(June), pp. 330–337. doi: 10.1016/j.socscimed.2018.06.038.

Maalouf, N. *et al.* (2018) 'Robotics in Nursing: A Scoping Review', *Journal of Nursing Scholarship*, 50(6), pp. 590–600. doi: 10.1111/jnu.12424.

MacIntyre, A. (2007) *After Virtue. A Study in Moral Theory*. 3rd ed. University of Notre Dame Press.

Van Maris, A. *et al.* (2020) 'The Impact of Affective Verbal Expressions in Social Robots', *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 508–510. doi: 10.1145/3371382.3378358.

Matarić, M. J. (2017) 'Socially assistive robotics: Human augmentation versus automation', *Science Robotics*, 2(4). doi: 10.1126/scirobotics.aam5410.

Matarić, M. J. and Scassellati, B. (2016) 'Socially Assistive Robotics', in Siciliano, B. and Khatib, O. (eds) *Springer Handbook of Robotics*. Springer, pp. 1973–1994.

McBride, N. (2020) 'Robot Enhanced Therapy for Autistic Children: An Ethical Analysis', *IEEE Technology and Society Magazine*. IEEE, 39(1), pp. 51–60. doi: 10.1109/MTS.2020.2967493.

Mejia, C. and Kajikawa, Y. (2017) 'Bibliometric Analysis of Social Robotics Research: Identifying Research Trends and Knowledgebase', *Applied Sciences (Switzerland)*, 7(12). doi: 10.3390/app7121316.

Miller, L. F. (2020) 'Human Rights of Users of Humanlike Care Automata', *Human Rights Review*. Human Rights Review, 21(2), pp. 181–205. doi: 10.1007/s12142-020-00581-2.

Misselhorn, C., Pompe, U. and Stapleton, M. (2013) 'Ethical Considerations Regarding the Use of Social Robots in the Fourth Age', *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry*, 26(2), pp. 121–133. doi: 10.1024/1662-9647/a000088.

Morozov, E. (2013) *To save everything, click here. The folly of technological solutionism*. Public Affairs.

Noori, F. M., Uddin, Z. and Torresen, J. (2019) 'Robot-Care for the Older People: Ethically Justified or Not?', *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE, pp. 43–47. doi: 10.1109/DEVLRN.2019.8850706.

Nylander, S., Ljungblad, S. and Jimenez Villareal, J. (2012) 'A complementing approach for identifying ethical issues in care robotics - Grounding ethics in practical use', *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*. IEEE, pp. 797–802. doi: 10.1109/ROMAN.2012.6343849.

O'Brolcháin, F. (2019) 'Robots and people with dementia: Unintended consequences and moral hazard', *Nursing Ethics*, 26(4), pp. 962–972. doi: 10.1177/0969733017742960.

Page, M. J. *et al.* (2021) 'The PRISMA 2020 statement: An updated guideline for reporting systematic reviews', *The BMJ*, 372. doi: 10.1136/bmj.n71.

Pareto Boada, J. (2021) 'Prolegómenos a una ética para la robótica social', *Dilemata, Revista Internacional de Éticas Aplicadas*, (34), pp. 71–87.

Payr, S. M. (2015) 'Towards Human-Robot Interaction Ethics', in Trappl, R. (ed.) *A Construction Manual for Robots' Ethical Systems. Cognitive Technologies*. Springer. doi: 10.1007/978-3-319-21548-8.

Pettit, P. (2002) *Republicanism. A Theory of Freedom and Government*. Oxford University Press.

Pranckutė, R. (2021) 'Web of science (Wos) and Scopus: The Titans of Bibliographic Information

in Today's Academic World', *Publications*, 9(12), pp. 1–59. doi: 10.3390/publications9010012.

Rabbitt, S. M., Kazdin, A. E. and Scassellati, B. (2015) 'Integrating socially assistive robotics into mental healthcare interventions: Applications and recommendations for expanded use', *Clinical Psychology Review*. Elsevier B.V., 35, pp. 35–46. doi: 10.1016/j.cpr.2014.07.001.

Richardson, K. (2019) 'The human relationship in the ethics of robotics: a call to Martin Buber's I and Thou', *AI and Society*. Springer London, 34(1), pp. 75–82. doi: 10.1007/s00146-017-0699-2.

Robson, A. (2018) 'Intelligent machines, care work and the nature of practical reasoning', *Nursing Ethics*, 26, pp. 1906–1916. doi: 10.1177/0969733018806348.

Rodogno, R. (2016) 'Social robots, fiction, and sentimentality', *Ethics and Information Technology*. Springer Netherlands, 18(4), pp. 257–268. doi: 10.1007/s10676-015-9371-z.

Schoeman, F. D. (ed.) (1984) *Philosophical Dimensions of Privacy: An Anthology*, *Philosophical Dimensions of Privacy*.

Sharkey, A. (2014) 'Robots and human dignity: A consideration of the effects of robot care on the dignity of older people', *Ethics and Information Technology*, 16(1), pp. 63–75. doi: 10.1007/s10676-014-9338-5.

Sharkey, A. and Sharkey, N. (2012) 'Granny and the robots: Ethical issues in robot care for the elderly', *Ethics and Information Technology*, 14(1), pp. 27–40. doi: 10.1007/s10676-010-9234-6.

Sorell, T. and Draper, H. (2017) 'Second thoughts about privacy, safety and deception', *Connection Science*, 29(3), pp. 217–222. doi: 10.1080/09540091.2017.1318826.

Sparrow, R. (2016) 'Robots in aged care: a dystopian future?', *AI and Society*. Springer London, 31(4), pp. 445–454. doi: 10.1007/s00146-015-0625-4.

Sparrow, R. (2019) 'Robotics Has a Race Problem', *Science, Technology, & Human Values*, p. 016224391986286. doi: 10.1177/0162243919862862.

Sparrow, R. and Sparrow, L. (2006) 'In the hands of machines? The future of aged care', *Minds and Machines*, 16(2), pp. 141–161. doi: 10.1007/s11023-006-9030-6.

Stahl, B. C. and Coeckelbergh, M. (2016) 'Ethics of healthcare robotics: Towards responsible research and innovation', *Robotics and Autonomous Systems*. Elsevier B.V., 86, pp. 152–161. doi: 10.1016/j.robot.2016.08.018.

Stokes, F. and Palmer, A. (2020) 'Artificial Intelligence and Robotics in Nursing: Ethics of Caring as a Guide to Dividing Tasks Between AI and Humans', *Nursing Philosophy*, (May), pp. 1–9. doi: 10.1111/nup.12306.

Tapus, A., Matarić, M. and Scassellati, B. (2007) 'The Grand Challenges in Socially Assistive Robotics', *IEEE Robotics and Automation Magazine*, 14(1), pp. 35–42.

The Member States (2012) *Charter of Fundamental Rights of the European Union*, *Official Journal of the European Union*. doi: 10.2307/j.ctt1ffjmjq.33.

Torralba, F. (2009) *La intimitat*. Pagès Editors.

Torras, C. (2019) 'Social networks and robot companions: Technology, ethics, and science fiction', *Metode*. Universitat de Valencia, 2019(9), pp. 163–169. doi: 10.7203/metode.9.12479.

Tzafestas, S. G. (2018) 'Roboethics: Fundamental concepts and future prospects', *Information (Switzerland)*, 9(6). doi: 10.3390/INFO9060148.

United Nations (2019) *World population prospects 2019. Highlights*, *Department of Economic and Social Affairs, Population Division*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12283219.

Vallès-Peris, N., Angulo, C. and Domènech, M. (2018) 'Children's Imaginaries of Human-Robot Interaction in Healthcare', *International Journal of Environmental Research and Public Health*. MDPI AG, 15(5). doi: 10.3390/ijerph15050970.

Vallès-Peris, N. and Domènech, M. (2020a) 'Roboticists' Imaginaries of Robots for Care: The Radical Imaginary as a Tool for an Ethical Discussion', *Engineering Studies*, 12(3), pp. 157–176. doi: 10.1080/19378629.2020.1821695.

Vallès-Peris, N. and Domènech, M. (2020b) 'ROBOTS PARA LOS CUIDADOS. LA ÉTICA DE LA ACCIÓN MESURADA FRENTE A LA INCERTIDUMBRE.', *Cuadernos de bioetica : revista oficial de la Asociacion Espanola de Bioetica y Etica Medica*, 31(101), pp. 87–100. doi: 10.30444/CB.54.

Vallor, S. (2015) 'Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character', *Philosophy and Technology*, 28(1), pp. 107–124. doi: 10.1007/s13347-014-0156-9.

Vallverdú, J. and Casacuberta, D. (2015) 'Ethical and technical aspects of emotions to create empathy in medical machines', in van Rysewyk, S. P. and Pontier, M. (eds) *Machine Medical Ethics*. Springer International Publishing, pp. 341–362. doi: 10.1007/978-3-319-08108-3_20.

Vandemeulebroucke, T., Casterle, B. D. and Gastmans, C. (2020) 'Ethics of socially assistive robots in aged-care settings: A socio-historical contextualisation', *Journal of Medical Ethics*, 46(2), pp. 128–136. doi: 10.1136/medethics-2019-105615.

Vandemeulebroucke, T., Dierckx de Casterlé, B. and Gastmans, C. (2018) 'The use of care robots in aged care: A systematic review of argument-based ethics literature', *Archives of Gerontology and Geriatrics*. Elsevier, 74(August 2017), pp. 15–25. doi: 10.1016/j.archger.2017.08.014.

Vanderelst, D. and Winfield, A. (2018) 'The Dark Side of Ethical Robots', *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, (December), pp. 317–322. doi: 10.1145/3278721.3278726.

Véliz, C. (2020) *Privacy is Power: Why and How You Should Take Back Control of Your Data*. Bantam Press.

Weng, Y. H. and Hirata, Y. (2018) 'Ethically Aligned Design for Assistive Robotics', in *2018 International Conference on Intelligence and Safety for Robotics*. IEEE, pp. 286–290. doi: 10.1109/IISR.2018.8535889.

van Wynsberghe, A. (2016) 'Service robots, care ethics, and design', *Ethics and Information Technology*. Springer Netherlands, 18(4), pp. 311–321. doi: 10.1007/s10676-016-9409-x.

van Wynsberghe, A. and Li, S. (2019) ‘A paradigm shift for robot ethics: from HRI to human–robot–system interaction (HRSI)’, *Medicolegal and Bioethics*, 9, pp. 11–21. doi: 10.2147/mb.s160348.

Yew, G. C. K. (2020) ‘Trust in and Ethical Design of Carebots: The Case for Ethics of Care’, *International Journal of Social Robotics*. Springer Netherlands, (April). doi: 10.1007/s12369-020-00653-w.

Zardiashvili, L. and Fosch-Villaronga, E. (2020) ‘“Oh, Dignity too?” Said the Robot: Human Dignity as the Basis for the Governance of Robotics’, *Minds and Machines*. Springer Netherlands, 30(1), pp. 121–143. doi: 10.1007/s11023-019-09514-6.

3. Ethics for social robotics: A critical analysis*

Júlia Pareto Boada
Institut de Robòtica i Informàtica Industrial, CSIC-UPC
Llorens i Artigas 4-6, 08028 Barcelona, Spain
jpareto@iri.upc.edu
ORCID iD: 0000-0003-4879-8800

Begoña Román Maestre
Facultat de Filosofia, Universitat de Barcelona
Montalegre 6, 08001 Barcelona, Spain
broman@ub.edu
ORCID iD: 0000-0001-6090-0172

Carme Torras
Institut de Robòtica i Informàtica Industrial, CSIC-UPC
Llorens i Artigas 4-6, 08028 Barcelona, Spain
torras@iri.upc.edu
ORCID iD: 0000-0002-2933-398X

Abstract: Social robotics' development for the practice of care and European prospects to incorporate these AI-based systems in institutional healthcare contexts call for an urgent ethical reflection to (re)configurate our practical life according to human values and rights. Despite the growing attention to the ethical implications of social robotics, the current debate on one of its central branches, social assistive robotics (SAR), rests upon an impoverished ethical approach. This paper presents and examines some tendencies of this prevailing approach, which have been identified as a result of a critical literature review. Based on this analysis of a representative case of how ethical reflection is being led towards social robotics, some future research lines are outlined, which may help reframe and deepen in its ethical implications.

Keywords: Care, Ethics, HRI, Justice, Social robotics, Well-being

1. Introduction

As a technoscientific activity developing tools for specific fields of professional human activity, social robotics is a principal actor in the practical and conceptual (re)configuration of our life. Like all technoscientific advances, it modifies the margins of human action. Still, it does so in an unprecedented way by allowing us to "outsource" part of our agency to robots in human practices of a relational kind, such as care. Robots' capacity to interact with humans "interpersonally" (Breazeal, Takanishi and Kobayashi, 2008) places social robotics as a promising technological contribution to European institutional care practices, mainly regarding healthcare (European Commission, 2020)(Dolic, Castro and Moarcas, 2019). Several European research initiatives and pilot projects (Ajuntament de Barcelona, 2020)(Andriella, Torras and Alenyà, 2020) reveal significant prospects to incorporate social robots within professional contexts of (health)care, especially for assistance (Dolic, Castro and Moarcas, 2019). This scenario urges to engage in an ethical reflection that may contribute to normatively guiding social robotics' disruptive force already from the early and throughout all different stages of its growing development (Mejia and Kajikawa, 2017), with a view to the European ideal of a human-

* **The text of this section entirely corresponds to the following publication: Pareto Boada, J., Román Maestre, B. and Torras, C. (2022) 'Ethics for social robotics: A critical analysis', in *TRAITS Workshop Proceedings (arXiv:2206.08270) held in conjunction with Companion of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*. Springer Berlin Heidelberg, pp. 1284–1286. https://doi.org/10.48550/arXiv.2207.12555**

centered technology at the service of human rights and well-being (High-Level Expert Group on AI, 2019). Although the widespread proliferation of ethical discussion on social robotics shows an increasing awareness of this urge, there are some flaws in the predominant ethical approach; that is, in the perspective from which the ethical problems are identified, framed and addressed. This has to do with the newness of social robotics as a field of specific ethical attention (Tzafestas, 2018) and a lack of a clear conceptual framework from which to engage in this normative-oriented kind of thinking.

In this paper, we present some main tendencies of the ongoing ethical reflection on social assistive robotics (SAR) disclosed through a critical literature review (Pareto Boada, Román Maestre and Torras, 2021), and which indicate significant shortcomings of the prevailing ethical approach. By discussing these tendencies, we aim to lay some theoretical grounds for a comprehensive ethical approach to social robotics in general, and to point at some new ethics research directions regarding its development for (institutional) care practices.

2. The ethical debate on SAR: three tendencies

As identified through previous work (Pareto Boada, Román Maestre and Torras, 2021), there are three significant tendencies of the current ethical reflection on SAR that should be redressed to enrich the debate on the implications of this technoscientific field of activity. In the following subsections, we set them out and briefly argue why they entail an impoverished ethical approach.

2.1. An individual-centered perspective

The ongoing ethical reflection on SAR is predominantly led from an individual-centered perspective, which focuses on the implications that robots may have for the well-being of humans with whom they interact. Much limited to the sphere of human-robot interaction (HRI), which is furthermore inadequately comprehended in dyadic terms (van Wynsberghe and Li, 2019)(Vallès-Peris, 2021), this ethical approach comes along with less attention to SAR implications from both the perspective of the specific (care) practice in which AI systems are introduced and the sociopolitical perspective of justice. This tendency means an important deficiency for a proper ethical approach, since it overlooks the constitutive interrelation between individual Well-being, Care and Justice as the main spheres of human activity with ethical importance regarding SAR. Indeed, an excessively restricted ethical focus at the individual level of human life and framed on the dyadic interaction between humans and robots implies overlooking the role of sociopolitical structure in the configuration of care practices, and thus the influence that these both have regarding individual Well-being. Thereby, this tendency unjustifiably falls in the "neglect of the political" underlying the mainstream philosophy of technology and ethics of technology, which some authors have already condemned and contributed to redressing (Coeckelbergh, 2018).

2.2. A narrowed understanding of ethical concepts

The individual-centered perspective seems to be tightly correlated to a second significant tendency of the ethical debate: a narrowed understanding of certain core ethical concepts around which SAR's problems stand, such as freedom and its related concept of autonomy, as well as responsibility. (Human) freedom is currently much restricted to what is philosophically

known as "negative liberty". Philosophical accounts of freedom (Berlin, 1969)(Pettit, 2002) and autonomy (Marzano, 2009) have, though, a richer scope of meaning, which enables a transversal gaze to SAR implications at the *micro*, *meso* and *macro* level of human life. The same happens with responsibility, which is currently understood in the traditional sense of liability for harm; a harm which is moreover linked to the robot's behavior. Philosophical approaches to responsibility offer a broader and more "substantial" understanding of the notion (Jonas, 2015), according to which the ethical approach would be also framed in terms of accountability for technological development (Coeckelbergh, 2021), thereby bringing to the fore discussion on the teleology and interests to which it is linked (perspective of justice). Therefore, the current restricted understanding of these notions comes along with an impoverishment of the ethical reflection.

2.3. A lack of discussion on SAR teleology

A third tendency of the current ethical approach is a general lack of explicit discussion on SAR teleology, that is, the "ends" at which it aims to serve, the "what for" of its development. Conceptual assumptions on "care" and "assistance", as well as other correlated notions that underlie the field's development (human well-being, human capabilities, autonomy) are usually not openly examined and revised, although some authors have already evinced the need of and engaged in such analysis (Aparicio Payá *et al.*, 2019)(van Wynsberghe, 2013). This means an important deficit in the predominant ethical reflection on SAR. The reason is to be found in the instrumental nature of technology, that is, in the fact that technology ultimately has an end that is external to itself, in the sense that its goal is to serve the purposes of the activities for which it is conceived as a means of support. This demands to conduct ethical reflection on SAR primarily as an exercise of applied ethics (Pareto Boada, 2021), understood as critical hermeneutics of human activity (Cortina, 1996). That is, the ethical implications of SAR must primarily be thought in the light of the specific practice for which technology is conceived, within the framework of its defining goals and values. For instance, the ethical issues of HRI must be examined in the light of the particular practice within which this interaction takes place: the conflicts of value coming into play in HRI will not be the same whether this interaction happens in the framework of a service activity (reception in a hotel) or of a care practice, such as when the interaction is a means to provide assistance in cognitive rehabilitation or company in front of solitude. Since SAR aims at contributing to practices of care, the scarce discussion on the constellation of SAR teleological-related meanings, altogether with the predominant ethical individual-centered perspective (instead of an approach primarily focused on the implications of SAR from the care-practice perspective) accounts for an impoverished ethical reflection.

3. Social robotics for care: towards a comprehensive ethical approach

Social robotics' development for care practices requires a comprehensive ethical approach that identifies and analyses the implications of this technoscientific field of activity at different levels of human life. As contended, this means, primarily, an exercise of applied ethics –in which reflection is contextualized according to social robotics' practical field of application–, coming along with a critical theory perspective –through which not only means, but also ends, are included in the discussion (which type of practices, societies and lives are to be reconfigured through social robotics' development?)–. Thus, an exhaustive ethical approach also demands to address social robotics' development from the macro perspective of justice. All this requires an

ethical gaze that takes into account conceptual advances of other subdisciplines of philosophy besides ethics, such as the refutations of the technology's value-neutrality thesis coming from the philosophy of technology (Heidegger, 1954) (Verbeek, 2006) (Verbeek, 2015), as well as the disclosure of the political dimension of technology set forth by political philosophy (Winner, 1980).

To achieve such a broader and deeper normative thinking on the ethical implications of social robotics for (institutional) care practices, we outline two main research lines to be next developed, considering the analyzed state of the art.

3.1. Approaching social robotics from a philosophical account of freedom and autonomy

Unfolding the philosophical concept of freedom and autonomy and ethically (re)examining social robotics in the light of these will broaden the ethical implications of the latter, by means of bringing to the fore the interdependence between freedom and the sociopolitical structuring of human life, as well as the political-structural dimension of human-technology relations. Issues concerning domination, manipulation or increased vulnerability raised by social robotics will appear not only regarding the interpersonal level of human life (linked to HRI in traditional terms) but also the structural one. This line of research could (partially) contribute to redressing the current individual-centered perspective.

3.2. Examining social robotics from the idea of "care"

Reflecting upon the notion of care as a practice –mainly drawing from J. Tronto's work (Tronto, 1993)– and identifying the ethical considerations regarding social robotics' development that emerge from this point of view will contribute to a proper exercise of (applied) ethics, by suitably redressing the primary focus of ethical attention to the sphere of the practice and enabling to filter reflection in the light of it. Moreover, analyzing the idea of care will help clarify and discuss the (shared) grounds of many ethical issues associated to HRI in assistive contexts (such as deception, dignity, emotional attachment, unauthentic intersubjectivity, objectification and recognition) (Pareto Boada, Román Maestre and Torras, 2021), as well as the technoscientific (Vallès-Peris and Domènech, 2020) and institutional corresponding narratives on innovation and implementation of social robots for care (European Parliament, 2017). By disclosing what is potentially at stake with the introduction of social robots in care practices, this analysis will help delineate the sort of ethical considerations that should be reflected upon when developing this technology for institutional practices of care.

ACKNOWLEDGMENTS

This work has been partially supported by the Spanish Ministry of Science and Innovation under a FPI scholarship for predoctoral contracts for the training of doctors (PRE2018-084286), and by the European Union Horizon 2020 Programme under grant agreement no. 741930 (CLOTHILDE).

REFERENCES

Ajuntament de Barcelona (2020) *Misty II the social robot becomes part of the lives of twenty senior citizens*. Available at: https://www.barcelona.cat/infobarcelona/en/tema/senior-

citizens/misty-ii-the-social-robot-becomes-part-of-the-lives-of-twenty-senior-citizens_907645.html (Accessed: 31 July 2021).

Andriella, A., Torras, C. and Alenyà, G. (2020) 'Cognitive System Framework for Brain-Training Exercise Based on Human-Robot Interaction', *Cognitive Computation*. doi: 10.1007/s12559-019-09696-2.

Aparicio Payá, M. *et al.* (2019) 'Un marco ético-político para la robótica asistencial. An Ethical-Political Framework for Assistive Robotics', *ArtefaCTos. Revista de estudios de la ciencia y la tecnología*, 8(1), pp. 97–117.

Berlin, I. (1969) *Four Essays on Liberty*. Oxford University Press.

Breazeal, C., Takanishi, A. and Kobayashi, T. (2008) 'Social Robots that Interact with People', in Siciliano, B. and Khatib, O. (eds) *Springer Handbook of Robotics*. Springer, pp. 1349–1369. doi: https://doi.org/10.1007/978-3-540-30301-5_59.

Coeckelbergh, M. (2018) 'Technology and the good society: A polemical essay on social ontology, political principles, and responsibility for technology', *Technology in Society*. Elsevier Ltd, 52, pp. 4–9. doi: 10.1016/j.techsoc.2016.12.002.

Coeckelbergh, M. (2021) 'Narrative responsibility and artificial intelligence', *AI & SOCIETY*. Springer London, (0123456789). doi: 10.1007/s00146-021-01375-x.

Cortina, A. (1996) 'El estatuto de la ética aplicada. Hermenéutica crítica de las actividades humanas', *Isegoría*, 13, pp. 119–134.

Dolic, Z., Castro, R. and Moarcas, R. (2019) *Robots in healthcare: a solution or a problem?, Study for the Committee on Environment, Public Health, and Food Safety*, *European Parliament*.

European Commission (2020) *White Paper on Artificial Intelligence - A European approach to excellence and trust*. doi: 10.1017/CBO9781107415324.004.

European Parliament (2017) 'Civil Law Rules on Robotics European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)', (July 1985), p. 23. Available at: http://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.pdf.

Heidegger, M. (1954) 'La pregunta por la técnica', *Revista de filosofía*, pp. 34–41.

High-Level Expert Group on AI (2019) 'Ethics Guidelines for Trustworthy AI'. European Commission, pp. 1–41.

Jonas, H. (2015) *El principio de responsabilidad. Ensayo de una ética para la civilización tecnológica.* Herder. Barcelona.

Marzano, M. (2009) *Consiento, luego existo. Ética de la autonomía*, *Proteus*. Proteus.

Mejia, C. and Kajikawa, Y. (2017) 'Bibliometric Analysis of Social Robotics Research: Identifying Research Trends and Knowledgebase', *Applied Sciences (Switzerland)*, 7(12). doi: 10.3390/app7121316.

Pareto Boada, J. (2021) 'Prolegómenos a una ética para la robótica social', *Dilemata, Revista*

Internacional de Éticas Aplicadas, (34), pp. 71–87.

Pareto Boada, J., Román Maestre, B. and Torras, C. (2021) 'The ethical issues of social assistive robotics: A critical literature review', *Technology in Society*, 67. doi: 10.1016/j.techsoc.2021.101726.

Pettit, P. (2002) *Republicanism. A Theory of Freedom and Government*. Oxford University Press.

Tronto, J. (1993) *Moral Boundaries. A Political Argument for an Ethic of Care*. Routledge.

Tzafestas, S. G. (2018) 'Roboethics: Fundamental concepts and future prospects', *Information (Switzerland)*, 9(6). doi: 10.3390/INFO9060148.

Vallès-Peris, N. (2021) 'Repensar la robótica y la inteligencia artificial desde la ética de los cuidados', *Teknokultura. Revista de Cultura Digital y Movimientos Sociales*, 18(2), pp. 137–146. doi: 10.5209/tekn.73983.

Vallès-Peris, N. and Domènech, M. (2020) 'Roboticists' Imaginaries of Robots for Care: The Radical Imaginary as a Tool for an Ethical Discussion', *Engineering Studies*, 12(3), pp. 157–176. doi: 10.1080/19378629.2020.1821695.

Verbeek, P.-P. (2006) 'Materializing Morality. Design Ethics and Technological Mediation', *Science, Technology, & Human Values*, 31(3), pp. 361–380.

Verbeek, P.-P. (2015) 'Beyond Interaction: A Short Introduction to Mediation Theory', *Interactions*, 22(3), pp. 26–31. doi: 10.1145/2751314.

Winner, L. (1980) 'Do Artifacts Have Politics?', *Daedalus*, 109(1), pp. 121–136. Available at: http://www.jstor.org/stable/20024652.

van Wynsberghe, A. (2013) 'Designing Robots for Care: Care Centered Value-Sensitive Design', *Science and Engineering Ethics*, 19(2), pp. 407–433. doi: 10.1007/s11948-011-9343-6.

van Wynsberghe, A. and Li, S. (2019) 'A paradigm shift for robot ethics: from HRI to human–robot–system interaction (HRSI)', *Medicolegal and Bioethics*, 9, pp. 11–21. doi: 10.2147/mb.s160348.

4. Social assistive robotics: An ethical and political inquiry through the lens of freedom*

Júlia Pareto Boada

Institut de Robòtica i Informàtica Industrial, CSIC-UPC

Llorens i Artigas 4-6, 08028 Barcelona, Spain

jpareto@iri.upc.edu
ORCID iD: 0000-0003-4879-8800

Mark Coeckelbergh

Department of Philosophy, University of Vienna

Universitätstrasse 7, 1010 Vienna, Austria

mark.coeckelbergh@univie.ac.at
ORCID iD: 0000-0001-9576-1002

Abstract: The development of social assistive robots for supporting institutional healthcare provision faces a lack of an ethical approach that adequately addresses the normatively relevant challenges regarding its deployment. Current ethical reflection is primarily informed by an individual-centered perspective focused on robots' implications for their end-users and thereby limited to the dyadic human-robot interaction sphere. Considering that this is tightly correlated to the restricted understanding of core ethical concepts upon which reflection stands, this paper delves into the concept of freedom from a philosophical perspective to unfold its full normative breadth for a critical assessment of technology's development. By bringing to the fore the political-structural dimension of freedom and, in turn, elaborating the political dimension of technology, the undertaken philosophical approach discloses freedom as a transversal ethical concept for a normative reflection on technology. Thereby, it broadens the scope of ethical attention beyond the sphere of human-robot interaction and turns attention to the so far overlooked structural dimension of human-robot relations. Using such philosophical idea of freedom, the paper approaches social assistive robotics and reexamines the terrain of relevant issues for its development. Since freedom is one major issue upon which current concerns revolve, the undertaken analysis substantially enriches the ongoing ethical discussion on social assistive robotics' implications for human freedom. In this way, this work contributes to going beyond the current individual-centered ethical perspective by laying conceptual grounds for a comprehensive ethical approach to social assistive robotics' development.

Keywords: ethics, freedom, healthcare, human-technology relations, political philosophy, social assistive robotics

1. Introduction

The rising development of social assistive robotics (SAR) (WIPO, 2021) as a tool-provider especially for the health care sector poses one of the new calls faced by contemporary ethics. As embodied AI systems that "socially" interact with humans as a means to carry out specific tasks, social assistive robots (SARs) are envisaged in Europe as a resource for professional activities of assistance (Dolic, Castro and Moarcas, 2019). That is, as tools for supporting tasks related to aiding people with special needs in different activities of their daily lives, which are held as care delivery either in institutional settings or at homes –e.g., physical and cognitive aid, capacities rehabilitation or maintenance, or even social needs management (Chita-Tegmark and Scheutz, 2021)–. Despite the outbreak of ethical attention and discussion on SAR that has come

*** The text of this section entirely corresponds to the original draft of the forthcoming publication: Pareto, J., Coeckelbergh, M. 'Social assistive robotics: An ethical and political inquiry through the lens of freedom'. Currently under review for its publication in the *International Journal of Social Robotics*.**

along with this scenario (Vandemeulebroucke, Casterle and Gastmans, 2020), the ethical call remains unanswered. The reason is that the predominant ethical approach to SAR is far from the one required for a normative-oriented thinking on technology and its deployment (Pareto Boada, Román Maestre and Torras, 2022).

As disclosed through previous work (Pareto Boada, Román Maestre and Torras, 2021), the ethical perspective from which SAR is primary being approached is an individual-centered one, which focuses on the implications that robots may have for the well-being of humans, specifically for their end-users. This is a narrow perspective, which informs an ethical reflection that is not only mainly engaged with technology implications at a *micro* level of human life, but that is moreover much limited to the sphere of a human-robot interaction (HRI) understood in dyadic terms. Thus, it is the individual life of SARs' interactants that is at the focus of ethical attention. Given the prevalence of this individual-centered perspective, the dominant ethical approach lacks of a due attention to SAR implications from both the perspective of the care practice in which SARs are (to be) deployed and from the perspective of justice. This is an important shortcoming that renders deficient any ethical approach to technology, and hence to SAR. There are several reasons for this.

First, it implies to overlook the constitutive interrelation between individual well-being, care practices and the sociopolitical activity and structure[43]; in other words, the interconnectedness between the *micro*, *meso* and *macro* levels of human life. By excessively restricting the focus of ethical attention at the individual level –moreover framing it on the dyadic interaction between humans and robots–, the latter is taken as quite disconnected from the kind of care practices and sociopolitical activity that indeed frame the conditions in which individuals' life takes place, and which thereby have a decisive role on their well-being.

Second –and this is related to the former to a great extent–, it entails to disregard the role of technology in the configuration of that political frame. This is what has already been reported as a neglect of the political dimension of technology (Coeckelbergh, 2022). That is, of the role that technology plays at the sociopolitical level, not only as an instrument for certain (disputable) ends, but also in the very same shaping of the conditions and structures within which life takes place. Such neglect is typical of an ethics of technology that is either worryingly disengaged from philosophy of technology, or bound to strands of this discipline that may fall short when it comes to attend the political and structural dimension of human-technology relations. In turn, this explains the undue disregard that political philosophy receives within the mainstream normative reflection on technology, where conceptual resources of this branch of philosophy are left unused, despite being fundamental for a proper framing of the normative questions raised by technoscientific activities (Coeckelbergh, 2018; 2022).

Alongside the individual-centered perspective, there is also a restricted understanding of some of the core ethical and political concepts from which SAR is normatively addressed, such as

[43] In this paper, the term "(sociopolitical) structure" is used to broadly refer to the institutional background of individuals' relational standing and activity in society (i.e., the political, legal, social and economic order of a society). In this sense, it is a term committed to the Rawlsian concept of "basic structure of society", defined as "[society's] main political, social and economic institutions and the how they fit together into one unified system of social cooperation from one generation to the next" and defended as the first subject of justice (Rawls, 2005, 11).

freedom and responsibility (Pareto Boada, Román Maestre and Torras, 2021). This is a conceptual limitation clearly correlated to the *micro* ethical perspective underlying the current approach to SAR, arguably in a causal sense. Indeed, the narrow way in which such notions are understood impoverishes the scope of issues that are normatively relevant regarding SAR development, and which should therefore be approached by a proper ethics of technology that takes into account the political dimension.

In response to these shortcomings of the predominant ethical approach to SAR, this paper unfolds the philosophical notion of freedom and uses it as a conceptual resource from which to ethically (re)examine this specific field of intelligent robotics in a comprehensive way that includes the political dimension.

The paper is structured as follows. The next section delves into the concept of freedom from a philosophical perspective; that is, by considering different accounts that have been developed within this discipline and which point at different dimensions of freedom that are relevant for an ethical reflection about, and an evaluation of SAR technology. Here, special attention is granted to the branch of political philosophy, since it provides key insights in this sense. Drawing upon certain contributions of philosophy of technology, section three then tackles the connection between technology and the *macro* level of human life, thereby examining the scope of human-technology relations beyond its "interpersonal" dimension[44]. This is undertaken as a complementary theoretical groundwork necessary for disclosing freedom as a philosophical notion that leads to a transversal ethical gaze to SAR implications at the three interrelated *micro*, *meso* and *macro* levels of human life. Section four then engages in the specific ethical analysis of SAR in the light of the unfolded understanding of freedom. Through it, we map the questions that are normatively relevant for SAR development, and show the limitations of current ethical reflection on this theme. In relation to this latter point, section four starts by reviewing the landscape of concerns on SAR that revolve around freedom, to enable contrasting it to that one informed by the full normative breadth of a philosophical account of this notion. To conclude, section five exposes how the undertaken work is a substantive contribution not only to the ongoing discussion on SAR and freedom, but also to a more comprehensive ethics for SAR, insofar as it redresses the individual-centered perspective towards one addressing SAR development in terms of justice.

2. The political-philosophical concept of freedom

As approached by philosophy, freedom is a concept with a rich and full of nuances scope of meaning, which gathers different constitutive dimensions of this human condition. Thus, turning to philosophy to delve into this idea provides a good ground for any normative reflection seeking to use freedom as analytic lens, as we here intend to.

[44] As explained in Sec. 4, this dimension refers to the kind of human involvement with technology that is defined by a direct interaction with or use of a technological artifact or system. The adjectivation "interpersonal" thus is not meant to be literally understood as refering to a relation between two persons, but, at most, between two agents –agency being then an unrefined concept dissociated from morality–. That is, if at all, it is to be taken in the minimalist terms of an "interagentive" relation.

To unfold the complex and multilayered philosophical idea of freedom, it is important to first recall I. Berlin's distinction between two different senses of this notion, namely "negative" and "positive freedom" (Berlin, 1969).

On the one hand, "negative freedom" refers to the absence of interference, obstacles, or constraints from others regarding one's own activity. This conception captures an essential dimension of freedom, which has to do with the fact of not being prevented from doing what an agent could do if no one would prevent them to. It is "freedom from" (Berlin, 1969), what it is being seized by this notion. Under it, freedom is to be understood as the extent of uncoerced[45] activity that is available to an agent.

On the other hand, "positive freedom" refers to the presence of control over one's own decisions and life, that is, to the exercise of self-determination. In this sense, freedom has to do with being the source of the reasons of one's own actions, in opposition to the fact of being subjected to others' reasons or to external causes as triggers of one's doing. A nice formulation of this account is to define freedom as the capacity to choose the reasons for which I choose (Román Maestre, 2021). Described as "freedom to" (Berlin, 1969), positive freedom is thus connected to the philosophical notion of autonomy as the capacity to govern oneself.

Accordingly, taken in its positive sense, freedom is infringed by phenomena like manipulation and paternalism –either be it a classical or a libertarian one (Coeckelbergh, 2022)–. As an intentional form of immorally influencing the other's decision-making in a certain direction, manipulation offends against their capacity to act as a (rational) self-directed agent. The same happens with paternalism, for it consists in deciding for another what is in their best interest – and therefrom interfering with their activity against their will (classical paternalism) or attempting to steer their choices towards what is taken to be their best interests (libertarian paternalism), as it is the case of nudging (Thaler and Sunstein, 2009)–.

The positive conception resonates well with a conceptualization of freedom from the so-called "capability approach" (Nussbaum, 2012). The idea of freedom as self-government implies that being free means something more than just not being interfered with one's own activity. Beyond this, it arguably entails to have certain abilities for self-determination –i.e., for reflecting upon reasons and purposes of one's own–, and also for effectively taking those as the basis for action[46]. From a positive account, then, the conditions of possibility of the exercise of autonomy are included under the scope of what being free means. In this sense, as a capacity (to govern oneself), freedom is best conceptualized in Nussbaumian terms of "capability"[47]. That is, as an

[45] Under this conception, impediments to freedom are understood as deliberate interferences, rather than mere contingencies beyond (direct) human control. Thus, uncoerced activity refers here to that activity that one can undertake insofar as it is not obstructed by other people, and not by factors such as natural causes or physical or mental limitations.

[46] It would be unlikely to talk about an agent being free were not possible for them to act according with their rational deliberation on the grounds for that action. Such conception would entail to reduce freedom to a mere exercise of reason, and to a condition proper of a *noumenical* subject. In turn, this would leave paternalism outside the scope of what counts as an infringement to freedom as autonomy, had the affected individuals exercised an own rational deliberation on the reasons for action –nonetheless disregarded as a valuable basis for a *de facto* self-government of these agents.

[47] Notice that therewith we do not mistakenly mean "freedom" to be a "capability" according to Nussbaum. Indeed, freedom is not included as such in her list of the ten central human capabilities.

actual capacity that, as such, is constitutively intertwined with the sociopolitical conditions that make it be a substantive rather than a formal faculty of self-definition –i.e., that make it be possibly exercised or materialized into a corresponding "functioning" (Nussbaum, 2012)–. In light of this, the idea of freedom is sharpened in a fundamental sense, namely in what we could call its political dimension. We will return to this later.

For the moment, it suffices to point out that, under the capability perspective, a new kind of threat to positive freedom can be identified, namely: the phenomenon of "adaptative preferences" (Nussbaum, 2012), which refer to preferences that result from a situation of constraint regarding the opportunity to choose according to one's one purposes and reasons, and which are therefore preferences built upon an adaptation to such basic limitation.

Besides those featured by the negative (non-interference) and positive (self-mastery) conceptions, there is another key dimension of freedom, which is the one highlighted by the contemporary republican conception of "freedom as non-domination" (Pettit, 2002). Freedom here refers to the absence of domination by others, where domination is understood as power, meaning the actual capacity of agents to arbitrarily interfere in certain choices that another is in a position to do. That is, the capacity of agents to intentionally obstruct the other's choice situation (interference) at their discretion, disregarding the interests and opinions of those affected (arbitrary). In this republican sense, then, freedom has to do with not being subjected to a *potential* arbitrary interference.

"Freedom as non-domination" thereby differs from the negative conception of freedom in that it is not (primarily) the absence of a *de facto* interference, what determines the extent of one's freedom, but rather the enjoyment of a certain (relational) status in which one is not exposed to such coercive phenomenon. Indeed, domination can take place without interference, as it happens when an agent is permanently exposed to arbitrary interference due to its status within a certain social structure. To clarify this, it is useful to recall the relationship slave-master that Pettit takes as an exemplary case of domination (Pettit, 2002): whether the master does actually coerce the slave or not, the latter is unfree because of being exposed to the possibility of an arbitrary interference by part of the master. In turn, interference can occur without domination, as it is the case when the coercive act does not respond to the exercise of a structural power asymmetry.

The republican approach to freedom as a matter primarily of no subjection rather than of no determination –in other words, its focus on power (a *potentia*) instead of on interference (a fact)– leads to a refined understanding of freedom as a relational condition. Otherwise said, as a capacity that is defined by the network of intersubjective relations in which the agent is placed (Marzano, 2009), as stressed by philosophy feminist critiques of the traditional atomistic conception of autonomy, which have led to reconceptualize it as "relational autonomy" (Mackenzie and Stoljar, 2000). In turn, this points at freedom as inherently linked to the sociopolitical structures under which such (power) relations are framed –i.e., to what we could refer to as a structural dimension of freedom–.

Rather, we just aim at highlighting that freedom, as a capacity (to govern oneself), is best read from a Nussbaumian understanding of capacity as "capability".

In that respect, a deeper reading of freedom as absence of domination is yet possible if understanding domination not only in the liberal republicanism sense but also in a Žižek/Marxian one, closer to the concept of "objective violence" (Žižek, 2011). That is, in terms of "structural domination", and not only of "intersubjective domination"[48], as we shall name them.

Under the liberal republican view, domination is understood as possibly exercised only by agents (either personal or collective) to other agents, but not by systems or networks (Pettit, 2002). It is in this sense that the liberal republican account of freedom as non-domination rests upon an idea of intersubjective domination. This rules out sociopolitical dispositions such as the institutional, economic and symbolic/ideological order as agents of domination. Žižek's conception of objective violence, though, helps remedy this, since it discloses another form of domination which is not attributable to particular individuals but rather precisely to that structural order. Unlike that kind of violence perpetrated by particular identifiable individuals (subjective violence), objective violence is impersonal, in that it is not traceable to specific subjects and intentions, but it comes from the same symbolic and structural order on which (intersubjective) power relations are grounded. This definition, insofar as it introduces the notion of an impersonal power of structures, contributes to disclose domination as taking place in a structural and not only an intersubjective form. As we shall see in Section 4, such conceptualization of domination introduces important nuances for thinking about the political relevance of technology, as it places the socio-symbolic order (in its linguistic, ideological and structural form) within the scope of consideration of technological development.

The so far examined philosophical conceptions of freedom disclose a key feature of this human capacity, which is its political-structural dimension. From a philosophical perspective, far from being a matter of individual capacity or exercise, freedom is constitutively linked to the sociopolitical framework within which it is (to be) exercised. Meaning, there is an intrinsic relation between freedom and the sociopolitical activity and structure in which individuals' life takes place. Non-interference, self-determination (or autonomy) and non-domination –both in intersubjective (republican) and structural terms–: all these senses of freedom –which we take as constitutive of a comprehensive understanding of this idea– point at this relation.

Whereas this has become evident from the non-interference and non-domination perspective of freedom, a brief remark may be useful to clarify the pivotal role of political activity in its depth regarding positive freedom. As contended, freedom requires certain conditions that make it possible for an agent to reflect upon their own reasons and to lead their actions as a result of this exercise –thereby enabling to pursue autonomy as effective freedom–. This appeals to political activity in that –borrowing Nussbaum's terminology–, these conditions have to do with fostering "combined capabilities", which are an agent's set of real opportunities to choose and act –and so the set of feasible "functionings"[49] that they can achieve in different spheres of life–

[48] Such an account has also been labeled as dyadic domination (Hasan, 2021). However, we will refer to it as "intersubjective domination", since it may better account for the republican conception of domination as possibly held not only by personal (individual) but also corporate/collective agents.

[49] Whereas "capabilities" refer to the doings and beings that people can achieve, "functionings" refer to the doings and beings that are realisations of that "capabilities". As exposed later in Section 4, such Nussbaumian distinction between capabilities and functionings is crucial for normatively addressing social robotics development for assistive contexts.

(Nussbaum, 2012). Since this set results from a combination of both internal abilities[50] (which are developed in interaction with the environment) and the political, social and economic framework in which the functioning associated with those can be actually exercised, freedom is clearly linked to politics. Regarding those (combined) capabilities, it is worth reminding that Nussbaum defines a list of ten central ones that must be protected up to a certain threshold level for a decent human life. Whereas discussing Nussbaum's commitment to specific capabilities and undertaking an exhaustive analysis of SAR from her capability approach would be beyond the scope of this paper, three of these central capabilities are indeed key with a view to a SAR development committed to human freedom, namely: "practical reason", "affiliation" and "control over one's own environment".

The political-structural dimension of freedom disclosed by the philosophical approach is of paramount importance for a normative reflection on freedom and its associated challenges or threats. Given that freedom has a dependence on the world and basic sociopolitical structures in which it is exercised, reflecting upon freedom demands, in turn, to critically reflect upon those former. This expands the scope of reflection on freedom beyond the *micro* or individual level of human life, and places the *meso* and especially *macro* spheres of human activity under analysis. Thinking about freedom issues necessarily entails to reflect upon the capabilities, power relations and social structures that condition and frame the exercise of this human capacity. Therefore, concerns on freedom must involve attention to justice, not only in distributive but also in structural terms.

3. Yes, but... what does Technology have to do with it? On human-technology relations

Now, when it comes to reflecting upon technology from ethics, a conceptual issue arises which is key to determining the normative relevance of that political-structural dimension of freedom regarding technological development: What does technology have to do with such dimension of freedom? In which sense is there a connection between technology and the sociopolitical frame within which freedom is exercised? Which role does technology play regarding the sociopolitical structuring of human life that decisively conditions freedom's development and exercise? In brief, the question at issue here revolves around what can be conceptualized as the political dimension of technology.

The reason why clarifying this issue becomes essential is a matter of argumentative consistency. That freedom is constitutively interrelated with the sociopolitical framework within which human life takes place –as highlighted by a philosophical approach– does not necessarily lead to turn attention to such framework as part of an ethical approach to technology. Unless, technology does have a role in its configuration, as it has indeed been disclosed by certain strands of philosophy of technology[51].

[50] Although Nussbaum originally names them "internal capabilities", following (Robeyns, 2017) we have chosen to use this alternative terminology to overcome the interpretation problems ascribed to Nussbaum's one.

[51] Notice that, by extension, this role has implications for the broader ethics of technology along the same lines: any normative reflection on technology must always include attention to issues arising from the perspective of justice.

How is, though, that idea of a political dimension of technology exactly to be understood? In what sense can technology be said to take an active part in the sociopolitical arrangement of human life?

A first attempt to provide an answer would be to recall the instrumental character of technology. In its constitutive dimension of being a means for certain ends, technology may serve as an instrument for specific political purposes, either be it or not deliberately conceived for those in the first place, and, more interestingly, either be these purposes realized through the immediate use of artifacts or through their very same design features. To illustrate this latter way in which technologies are (instrumentally) political, it suffices to think about the well-known case of R. Moses' bridges over the parkways on Long Island. The intentionally-devised architectural characteristics of the bridges prevented low-income classes and racial minorities to reach Jones Beach (Winner, 2009), thereby imposing a certain social order characterized by inequality in terms of substantive opportunities. The instrumentality of technology, then, for it places artifacts as possible means of settling a specific order of social relationships within human collectives (Winner, 2009), provides a ground to talk about a political dimension of technology.

For the instrumentality of technology to be a proper ground to attribute a political dimension to technology, though, that instrumental character is not to be naïvely understood as morally neutral. Against the modern conception (Feenberg, 2018), technology is not, in itself, a value-free means for reaching certain ends. This is so for at least two reasons.

On the one hand, because as the product of a specific sociohistorical context, technology always entails a particular cosmovision (Hui, 2020). Technologies arise from a specific way of understanding and relating to the world and to others –in virtue of which these are precisely conceived as tools–. In this sense, insofar as it responds to a particular commitment to reality on the basis of anthropological, epistemic and political assumptions, technology is, from its very same conception, axiologically charged. Contemporary critical theory of technology has made a point of this by stressing the social character of technology, that is, the interconnection between social meanings (and relations) and functional rationality (Feenberg, 2009a) .

On the other hand, because technology is not only the product but also the source of a particular way of thinking about and being in the world, as stressed already from the classical phenomenological analysis of technology. Technology frames the way in which the world appears to us and thereby also conditions how we understand and relate to it (Heidegger, 2009). As it were, technologies open up the world in certain (rather than other) forms. Postphenomenology has refined this insight by pointing at the mediating role of technologies in human-world relationships (Rosenberger and Verbeek, 2015), both in hermeneutic and existential terms (Ihde, 2009)(Verbeek, 2005). Technological artifacts actively shape how reality appears to and is interpreted by us, as well as how we act and organize our lives. On the grounds of these two main dimensions of mediation –mediation of perception or experience and mediation of action or praxis–, technologies can be contended to shape moral decisions and, thus, to mediate morality too (Verbeek, 2011). Insofar as they provide access to reality in certain ways rather than others, and invite or inhibit certain actions rather than others (Verbeek, 2011), then, technologies are far from neutral instruments.

This phenomenon of "technological mediation" is relevant for an accurate account of the political dimension of technology. Although this –with the exception of some more recent attempts (Verbeek, 2020)– has been left quite unexplored by the very same postphenomenology (Feenberg, 2009b) (Kaplan, 2009) (Verbeek, 2009) (Coeckelbergh, 2020), understanding technologies as mediators of human-world relations opens up a further sense in which technology has a constitutive role at the sociopolitical level of human life, beyond that one attributable to its instrumental dimension. Ultimately, this has to do with a double level of technological mediation. Technologies' role in framing humans' hermeneutical and existential stance with and within the world applies both for the *micro* and *macro* level of human life. As pointed at by latest postphenomenological studies, technologies do not only influence the user's perception and experience, and thereby their decisions and actions, but they also have a constitutive role in cultural and moral frameworks of interpretation, and in social practices (Kudina and Verbeek, 2019) (Verbeek, 2020). It is in this sense that, from a postphenomenological perspective, human-technology relations are contended to have a political dimension (Verbeek, 2020), which goes beyond the more strictly instrumental dimension of such relations –that is, beyond the politics that these relations may entail in virtue of the instrumental use of artifacts towards certain ends–. Despite a more specific analysis on the praxis side of technological mediation at the *macro* level would still be in order to thoroughly delve into this, the connection between technological mediation and the politics of technology seems undeniable. By shaping human experiences, margins of action and situations of choice, specific social and thus power relations are also framed and organized through technologies. Consider for example birth control pills, which, by disconnecting sex from reproduction, changed women' and also homosexuals' relational standing within society. But also SARs are likely to change power relations, as we will show below. This connects to critical theory's understanding of technologies as frameworks for ways of life, rather than as mere tools (Feenberg, 2018).

Notice that technological mediation also allows to make better sense of the phenomenon of the non-intentional political implications of technology, so well highlighted by Winner through different case examples (Winner, 2009). That is, the fact that technology has a configurating role in the social order beyond that which may be intentionally sought –i.e., beyond the one that it has in virtue of being purposedly used as an instrument for arranging power and authority –.

Drawing upon contemporary philosophical insights, then, it is not only as (non-neutral) instruments that may serve specific political ends and distribution of power, but also in virtue of their mediating character, that technologies do have a political dimension.

Such account of the politics of technology refines the understanding of human-technology relations in a way that is key for a normative reflection and technological development, for it discloses what we will call the structural dimension of such human-technology relations. Beyond an "interpersonal" level of human-technology relations, in which humans relate with (mediating) technologies in terms of use or of interaction with artifacts, the relations between humans and technologies are also of a structural kind. That is, technologies do enter in relation with humans –thereby having implications for human life– not only in virtue of human direct involvement (usage or interaction) with them, but also more indirectly in virtue of their active coshaping of frameworks for life. Thus, humans relate with technologies not merely as users or interactants with those, but as subjects within technologically coshaped social structures. In next

section, we will show the extent to which this dimension calls for a reconsideration of the scope of current ethical discussion on the specific technoscientific field of SAR.

4. (Re)Examining SAR in the light of freedom

In the light of the philosophical idea of freedom, which accounts for the connection between this capacity and the sociopolitical framework of human life, and taking into account the configurating role that technology plays at that macro level (i.e., the political dimension of technology), let us now turn specific attention to social assistive robotics (SAR) to (re)examine the normatively relevant issues that arise regarding its development. Recent illustrative instances of SARs are the socially interactive AI robotic system that helps dementia patients in cognitive training exercises (Andriella, Torras and Alenyà, 2019) (Andriella, Torras and Alenyà, 2020); and the social assistive robot Misty II that helps dependent elderly that live alone with their needs in domestic daily life (Ajuntament de Barcelona, 2020).

A previous contextualization is in order. As identified through a critical literature review (Pareto Boada, Román Maestre and Torras, 2021), in the landscape of scholar ethical reflection on SAR, freedom –referred to either (and sometimes indistinctly) as human "freedom" or "autonomy"– is one of the main issues upon which concerns do revolve. However, the mainstream understanding of freedom in current ethical discussion proves to be a quite narrow one, in that it overlooks some significant dimension of this human capacity. Especially, the political-structural dimension stressed by philosophical conceptualizations, as a revision of current freedom-related concerns shows. So, whereas approaching SAR challenges from the perspective of freedom is not a novel move –hence that we speak of a *re*examination–, doing it from the philosophical notion of freedom is indeed, since it maps a new terrain of freedom-related issues that so far remain out of the scope of ethical consideration.

So far, the problems that SAR raises for human freedom are mainly being thought in individual key. They mostly concern the hampering of individuals' exercise of this agency-related faculty, specifically that of robots' end-users. Indeed, were we to summarize the different ways in which SAR is considered to challenge freedom in current literature (Pareto Boada, Román Maestre and Torras, 2021), we could talk about two main kinds of endangerments to freedom associated with this branch of robotics. On the one hand, the interference with users' action and decision, basically in virtue of robots' technological autonomy –through which, even if it is in the name of the user's well-being and health, these AI systems may restrict humans' courses of action (classical paternalism case)–. On the other hand, the exploitation or fostering of human autonomy's vulnerabilities, in terms of users' capacity both for (functional) independence ("can I do 'x' or not by my own?") and (chiefly Kantian) self-determination ("for which reason do I choose to do 'x' or not?). Whereas challenges to independence relate to the risks of an inappropriate assistance that leads to a users' loss of capacities alongside a dependency on technology (e.g., the user was previously able to eat alone, but stopping exercising the functioning has weakened their capacity to do so), challenges to self-determination mostly relate to the risks of socially interacting with the robot, such as human exposure to manipulation, emotional attachment and improper decision-making delegation to the robot. Only on few occasions is SAR endangerment to self-determination argued in terms of the violation of users' capacity to live according to their own reasons that robots' implementation entails, inasmuch as it is grounded on interests alien to end-users' ones.

Such landscape of concerns thus reveals that problems for freedom are primarily and almost exclusively being set as problems for the freedom of individuals with whom robots interact, insofar as they are mostly taken as a matter of direct use or interaction with robots, i.e., as problems that arise within the context of (dyadic) human-robot interaction (HRI).

From a philosophical viewpoint, this is representative of a predominant micro perspective of freedom as an individual exercise of a human faculty, disconnected from the kind of practices, power relations and sociopolitical structures in which the latter is exercised –i.e., the *meso* and *macro* levels of human life–. Indeed, drawing upon the multilayered philosophical idea of freedom, current concerns appear to be (more or less exhaustively) informed by the two dimensions of freedom related to the negative and positive conceptions, namely: non-interference and self-government or autonomy. This explains that freedom issues relate to the fact that robots may coerce the courses of action of their users as well as infringe upon their ability to act as rational self-directed agents. What remains out of the scope of attention, though, is the political-structural dimension of freedom, so well stressed by a capability-approach version of positive freedom and by contemporary conceptualizations of freedom as non-domination, which account for the relationality of autonomy. And, with it, the political dimension of technology, which evinces the flaw of an exclusive focus on the "interpersonal" level of human-robot relations.

In the light of these other defining senses of freedom so far overlooked in ethical discussion, more issues appear as essentially related to SAR implications for human freedom. Let us now examine this, beginning with freedom as non-domination.

First, from the (liberal) republican understanding of freedom as a relational condition of no subjection to potential arbitrary interference, SAR implications for human freedom concern the kind of intersubjective (power) relations that are fostered by such technological development –on the grounds of which individuals' autonomy is actually delimited as a capability[52]–.

Of course, this implies the relations between those particular agents involved in the care practice in which robots are implemented. Given the asymmetry of power that care relationships entail, this is a key topic regarding the development of technologies for assistive contexts. SARs' introduction may reframe power between care-receivers and (formal or informal) caregivers in a way that raises the exposure to arbitrary interference –realizable either by robots or humans in charge– within care relations. For instance, the data gathering that SARs may undertake in their daily domestic assistance could expose elder end-users to paternalistic interferences by their caregivers (doctors or relatives) to a greater extent than before. Furthermore, SARs implementation may place new agents of power in former relational networks of care, which may bring with new modes of subjection to interests alien to the care practice defining ones. For instance, inasmuch as private technology companies become part of such networks, the power between the public healthcare sector and private companies may be rearranged in terms of the defining of healthcare models.

Thinking in these non-domination terms also applies to relations with and between "nonnurturant" (Duffy, 2007) care agents, i.e., workers that carry on the nonrelational tasks of care (otherwise called "dirty work"), such as cleaning, food preparation or service. SARs'

[52] See footnote 5.

introduction may sharpen existing instances of domination in labour relations around that part of care (e.g. rising "nonnurturant" workforce's vulnerability to exploitation). At the same time, it may bring with new professional categories of "nonnurturant" care (e.g., related to the development, functioning and maintenance of the robotic AI systems) that may also reshape power in such terms, even at a global level. For example, relations of exploitation between tech companies and workforces of data labeling for the training of AI systems (Crawford, 2021) may also arise from SAR development.

Notice that an approach to SAR in the light of the (liberal) republican account of non-domination thus points at a reflection in structural key. That is, it introduces the sociopolitical structure of intersubjective relations that robotics help configure within the scope of ethical consideration, thereby taking seriously the social dimensions of personal autonomy (Mackenzie, 2021). Thinking about relations implies attending them also in broader terms of structural networks: Which particular social order is maintained or created through SAR development and implementation?

Second, from a Žižek/Marxian understanding of freedom as non-domination, SAR implications for human freedom concern the socio-symbolical and structural order that SAR helps constitute –which ultimately grounds the relational standing of subjects–. In the light of structural domination, at least two main areas of consideration arise regarding SAR development: an ideological and a more infrastructural-related one, that is, having to do with the mechanisms of human functioning.

On the one hand, comprehending that the socio-symbolical order functions as an (impersonal) form of power –in that it molds situated subjectivities– demands to reflect upon the kind of anthropological and sociopolitical assumptions and narratives that SAR entails and fosters. Which conceptions of the relational selves does robotics promote or reinforce through, for instance, the design and the roles assigned to robots? SARs' appearance-related traits such as gender, colour or the AI-systems' voice (UNESCO, 2019), altogether with the role that these artifacts endorse, may strengthen social stereotypes, bias and inequality on morally irrelevant grounds such as gender or race. Which comprehensions of the meaningful life, vulnerability, autonomy and other care-related ideas are set through SAR development, thereby constituting the socio-symbolical background on which individuals stand in relation to themselves and others? For instance, the kind of assistance that SAR is conceived for, insofar as it is exclusively aimed at replacing individuals' (disfunctioning) abilities, may reinforce capacitism by sustaining the medical-rehabilitation model of disability to the detriment of the social one (Aparicio Payá *et al.*, 2019).

On the other hand, in the light of structural non-domination, freedom not only relates to the socio-symbolical order, but it also has to do with what we will here refer to as the infrastructure of human functioning. In this sense, domination has to do with the framing of human functioning into certain restrictive patterns. On one side, this framing may well come as a self-imposed regulation of one's own doing triggered by the kind of "panopticism" (Foucault, 1977) that SARs entail. Picking up on the issue of SARs' potential data gathering in domestic environments, for instance, not only end-users but also the people entering their private sphere (e.g, relatives or other visitors) may modify their behavior, conversation or interaction in the face of the surveillance under which SARs' use may put them. On the other side, though, structural

domination in this sense is a more even complex issue that directly links to the phenomenon of technological mediation or, otherwise raised, to the fact that technology constitutes frameworks for certain ways of life. In light of this, SAR implications for human freedom have to do with how this technology delimitates the margins of conversion of capacities to functionings. Which kind of structural order of functioning does this technology imply, and does it restrict human capabilities? Does SAR structurally frame capacities to do and be in a way that limits the range of available options rather than expands them? An enlightening question to land this matter: Are there alternative options to turn our capacities into functionings outside the conversion order that SAR technology provides? SARs' implementation in elder-care facilities to assist them in daily tasks such as dressing or eating could for instance configure an infrastructure of functioning that restricts the elderly's alternative ways of operating towards these ends[53].

In turn, this connects to the third sense of freedom left insufficiently addressed within the mainstream ethical discussion: positive freedom or autonomy understood in capability terms, that is, as an actual faculty to conduct one's life according to one's reasons and purposes. Under it, SAR implications for human freedom concern the set of opportunities for being and doing that robotics helps configure, not only in terms of individual capacities but also more structurally in terms of "functioning environments" (Aparicio *et al.*, 2020). Let us unpack this.

In the light of such account of freedom, the question is the extent to which SAR does stimulate, preserve or amplify human possibilities to be and do, and for whom. Within assistive contexts, this first relates to Kantian autonomy, thus concerning the capacity to choose the reasons for which one chooses (Román Maestre, 2021), in this case that of end-users. In this sense, a basic instance of freedom's infringement is SARs' use on the grounds of an "adaptative preference" (Nussbaum, 2012), as it would be the case had an elderly person decided to introduce this assistive AI technology at home not because this adjusts to their own reasons and purposes, but because of being this a preference resulting from a lack of alternative options at disposal ("Better this rather than nothing" would be then the underlying reasoning of such act of consent).

A brief remark is in order here: against what seems to be the mainstream tendency in current ethical reflection, understanding care properly as a relational practice could imply that not only end-users and healthcare professionals but also patients' relatives should be considered when reflecting upon SAR development in the light of positive freedom.

At the same time, committing to freedom as an (effective) capacity to self-determination demands SARs' development and implementation to serve certain Nussbaumian central human capabilities. Namely, the architectural capabilities of "practical reason" and "affiliation", as well as of "control over one's environment", insofar as they are essential to conduct one's life according to one's own reasons and purposes. In turn, this leads us to a crucial related issue: to distinguish well between capabilities and functionings[54] regarding the goal of assistance provision –and, thereby, what SARs should ultimately aim at supporting–. Indeed, with a view to freedom, there is a decisive distinction between enabling capabilities (and thus the

[53] This is in line with what has already happened, for example, with the infrastructure of functioning co-created by the extended use of mobile phones, outside of which we cannot easily function regarding significant facets of life anymore, such as work or even healthcare services' access.
[54] See footnote 7.

opportunity to exercise them or put them to function) and imposing functionings without fostering capabilities. For example, imagine that SARs were introduced in an eldercare facility to assist physically disabled residents to get dressed, but these end-users had no choice to decide why or why not (so when and in what circumstances) to use the robot –but instead, they were by default daily forced to rely on the robotic-assistance (e.g. for the sake of the timetable dynamics). In this case, it would be the end-users' functionings but not their capabilities, what SAR would foster. This would infringe upon freedom as (actual) autonomy, for the robots' use would neither respond to the user's capacity to reason on the reasons for action (practical reason), neither promote the opportunities to exercise it.

Sticking to distinction between capabilities and functionings, the questions on SAR development arising from the capability perspective are not only to be filtered in individual, but also in structural terms. The focus is here on the kind of "functioning environments" that SAR shapes, in the sense of the kind of infrastructural dispositions for converting individuals' capacities into functionings that are (co)created by SAR. Does SAR expand rather than restrict the ways of realizing diverse capacities, and thus the range of potential functionings? Notice that this renders the capability-issues intertwined with the second kind of structural domination-related issues stressed above, by linking the former to the conversion order or infrastructure of human functioning that technology brings with. Yet, it also points to a further and broader issue, in that it calls to reflect even broader upon the material socio-technical configuration that SAR provides, and about whether it is responsive to "social vulnerability" (Liedo, 2021). For instance, SAR focus on promoting personal autonomy by means of assisting in disabilities could overlook that the link of human autonomy to social vulnerability, and thus the role that assistive technologies could undertake regarding the provision of material facilitating environments for an egalitarian conversion of diverse capacities into functionings.

In sum, in the light of a philosophical understanding of freedom, SAR challenges for freedom concern not only the individual exercise of such an agency-related capacity, but also the basic sociopolitical structures that constitutively frame and condition its exercise, and which SAR helps configure. Thus, the implications of SAR development for healthcare concern the kind of intersubjective relations, power structures and human capabilities that this AI assistive robotic technology contributes to shape and foster. Consequently, the philosophical perspective substantively remaps the current terrain of normative consideration on SAR development with a view to freedom, broadening reflection towards SAR implications for the political-structural dimension of such human capacity. In turn, this brings to the fore of ethical concern the so far overlooked structural dimension of human-robot relations, thus properly addressing the political dimension of technology.

5. Conclusions

In response to the significant shortcomings of the predominant ethical approach to SAR, this paper has advanced the philosophical idea of freedom as a transversal ethical concept for the normative reflection on technology development. That is, it has been developed as a concept that leads to critically addressing SAR development in terms of its implications at the *micro*, *meso* and *macro* levels of human life, i.e., regarding its disruptive potential for the individual's life, the (health)care practices, and the sociopolitical structure.

Drawing upon relevant philosophical accounts that gather several constitutive dimensions of this human capacity, this paper has disclosed a feature of freedom that is key for an ethical reflection in the light of this notion, namely its political-structural dimension of freedom. That is, the fact that freedom is constitutively linked to the sociopolitical framework within which it is (to be) exercised, and thus, far from being an individual capacity or exercise, it is related to the sociopolitical structuring of human life. Revealing the proactive role of technology in the configuration of the sociopolitical framework in which individuals' life takes place, this paper has showed the pertinence of taking this political-structural dimension under the scope of ethical attention on SAR. Making use of the normative breadth of the unfolded philosophical notion, the relevant issues for SAR development with a view to freedom have been reexamined. In this way, this paper has substantially enriched the current discussion on SAR implications for human freedom by broadening the scope beyond the sphere of human-robot interaction (HRI) and bringing to the fore the so far overlooked structural dimension of human-robot relations.

Funding: This work has been partially supported by grant PRE2018-084286 funded by MCIN/ AEI/10.13039/501100011033 and by "ESF Investing in your future", and by the European Union Horizon 2020 Programme under grant agreement no. 741930 (CLOTHILDE).

Competing Interests: The authors have no competing interests to declare that are relevant to the content of this article.

Author's contributions: All authors contributed to this work. The first draft of the manuscrit was written by Júlia Pareto-Boada and all authors commented on and critically revised the previous versions of the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We want to express our gratitude to Dr. Begoña Román and Dr. Carme Torras for their fruitful discussions on this work, which substantially contributed to refine it.

References

Ajuntament de Barcelona (2020) *Misty II the social robot becomes part of the lives of twenty senior citizens*. Available at: https://www.barcelona.cat/infobarcelona/en/tema/senior-citizens/misty-ii-the-social-robot-becomes-part-of-the-lives-of-twenty-senior-citizens_907645.html (Accessed: 31 July 2021).

Andriella, A., Torras, C. and Alenyà, G. (2019) 'Short-Term Human–Robot Interaction Adaptability in Real-World Environments', *International Journal of Social Robotics*. Springer Netherlands. doi: 10.1007/s12369-019-00606-y.

Andriella, A., Torras, C. and Alenyà, G. (2020) 'Cognitive System Framework for Brain-Training Exercise Based on Human-Robot Interaction', *Cognitive Computation*. doi: 10.1007/s12559-019-09696-2.

Aparicio, M. *et al.* (2020) 'Discursive Frameworks for the Development of Inclusive Robotics', *Biosystems and Biorobotics*, 25, pp. 74–80. doi: 10.1007/978-3-030-24074-5_14.

Aparicio Payá, M. *et al.* (2019) 'Un marco ético-político para la robótica asistencial. An Ethical-Political Framework for Assistive Robotics', *ArtefaCTos. Revista de estudios de la ciencia y la tecnología*, 8(1), pp. 97–117.

Berlin, I. (1969) *Four Essays on Liberty*. Oxford University Press.

Chita-Tegmark, M. and Scheutz, M. (2021) 'Assistive Robots for the Social Management of Health: A Framework for Robot Design and Human–Robot Interaction Research', *International Journal of Social Robotics*. Springer Netherlands, 13(2), pp. 197–217. doi: 10.1007/s12369-020-00634-z.

Coeckelbergh, M. (2018) 'Technology and the good society: A polemical essay on social ontology, political principles, and responsibility for technology', *Technology in Society*. Elsevier Ltd, 52, pp. 4–9. doi: 10.1016/j.techsoc.2016.12.002.

Coeckelbergh, M. (2020) *Introduction to Philosophy of Technology*. Oxford University Press.

Coeckelbergh, M. (2022) *The Political Philosophy of AI* . Polity Press.

Crawford, K. (2021) *Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

Dolic, Z., Castro, R. and Moarcas, R. (2019) *Robots in healthcare: a solution or a problem?, Study for the Committee on Environment, Public Health, and Food Safety*, *European Parliament*.

Duffy, M. (2007) 'Doing the dirty work: Gender, race, and reproductive labor in historical perspective', *Gender and Society*, 21(3), pp. 313–336. doi: 10.1177/0891243207300764.

Feenberg, A. (2009a) 'Democratic Rationalization: Technology, Power, and Freedom', in Kaplan, D. M. (ed.) *Readings in the Philosophy of Technology*. 2nd ed. Rowman & Littlefield Publishers, Inc.

Feenberg, A. (2009b) 'Peter-Paul Verbeek: Review of What Things Do', *Human Studies*, 32(2), pp. 225–228. doi: 10.1007/s10746-009-9115-3.

Feenberg, A. (2018) 'What Is Philosophy of Technology?', in Beira, E. and Feenberg, A. (eds) *Tecnology, Modernity, and Democracy. Essays by Andrew Feenberg*. Rowman & Littlefield International.

Foucault, M. (1977) *Discipline and Punish: The Birth of the Prison*. New York: Vintage Books.

Hasan, R. (2021) 'Republicanism and Structural Domination', *Pacific Philosophical Quarterly*, 102(2), pp. 292–319. doi: 10.1111/papq.12337.

Heidegger, M. (2009) 'The Question Concerning Technology', in Kaplan, D. M. (ed.) *Readings in the Philosophy of Technology*. 2nd edn. Rowman & Littlefield Publishers, Inc., pp. 9–24.

Hui, Y. (2020) *Fragmentar el futuro: ensayos sobre tecnodiversidad*. Buenos Aires: Caja Negra.

Ihde, D. (2009) 'A Phenomenology of Technics', in Kaplan, D. M. (ed.) *Readings in the Philosophy of Technology*. 2nd Ed. Rowman & Littlefield Publishers, Inc.

Kaplan, D. M. (2009) 'What Things Still Don't Do', *Human Studies*, 32(2), pp. 229–240. doi: 10.1007/s10746-009-9116-2.

Kudina, O. and Verbeek, P. P. (2019) 'Ethics from Within: Google Glass, the Collingridge Dilemma, and the Mediated Value of Privacy', *Science Technology and Human Values*, 44(2), pp. 291–314. doi: 10.1177/0162243918793711.

Liedo, B. (2021) 'Vulnerabilidad', *Eunomía. Revista en Cultura de la Legalidad*, 20, pp. 242–257. doi: https://doi.org/10.20318/eunomia.2021.6074.

Mackenzie, C. (2021) 'Relational Autonomy', in Hall, K. Q. and Ásta (eds) *The Oxford Handbook of Feminist Philosophy*. Oxford Academic. doi: https://doi.org/10.1093/oxfordhb/9780190628925.013.29.

Mackenzie, C. and Stoljar, N. (eds) (2000) *Relational Autonomy: Feminist Perspectives on Autonomy, Agency and the Social Self*. Oxford University Press.

Marzano, M. (2009) *Consiento, luego existo. Ética de la autonomía*, *Proteus*. Proteus.

Nussbaum, M. C. (2012) *Crear capacidades. Propuesta para el desarrollo humano*. Paidós.

Pareto Boada, J., Román Maestre, B. and Torras, C. (2021) 'The ethical issues of social assistive robotics: A critical literature review', *Technology in Society*, 67. doi: 10.1016/j.techsoc.2021.101726.

Pareto Boada, J., Román Maestre, B. and Torras, C. (2022) 'Ethics for social robotics: A critical analysis', in *TRAITS Workshop Proceedings (arXiv:2206.08270) held in conjunction with Companion of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*. Springer Berlin Heidelberg, pp. 1284–1286.

Pettit, P. (2002) *Republicanism. A Theory of Freedom and Government*. Oxford University Press.

Rawls, J. (2005) *Political Liberalism*. Expanded E. New York: Columbia University Press.

Robeyns, I. (2017) 'Clarifications', in *Wellbeing, Freedom and Social Justice: The Capability Approach Re-Examined*. Open Books, pp. 89–168.

Román Maestre, B. (2021) 'Llibertat. Idees clàssiques pel món que ve. 1/3'. Available at: https://www.instituthumanitats.org/ca/documents/videos/1-3-idees-classiques-per-al-mon-que-ve-llibertat-sessio-1.

Rosenberger, R. and Verbeek, P.-P. (eds) (2015) *Postphenomenological Investigations: Essays on Human–Technology Relations*. Lexington Books.

Thaler, R. H. and Sunstein, C. R. (2009) *Nudge: Improving Decisions about Health, Wealth, and Happiness*. London: Penguin.

UNESCO (2019) *I'd blush if I could: Closing gender divides in digital skills through education*. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1.

Vandemeulebroucke, T., Casterle, B. D. and Gastmans, C. (2020) 'Ethics of socially assistive robots in aged-care settings: A socio-historical contextualisation', *Journal of Medical Ethics*, 46(2), pp. 128–136. doi: 10.1136/medethics-2019-105615.

Verbeek, P.-P. (2005) *What Things Do: Philosophical reflections on technology, agency, and design*. The Pennsylvania State University Press.

Verbeek, P.-P. (2009) 'Let's Make Things Better: A Reply to My Readers', *Human Studies*, 32(2), pp. 251–261. doi: 10.1007/s10746-009-9118-0.

Verbeek, P.-P. (2011) *Moralizing Technology: Understanding and Designing the Morality of Things*. The University of Chicago Press.

Verbeek, P. P. (2020) 'Politicizing Postphenomenology', in Miller, G. and Shew, A. (eds) *Reimagining Philosophy and Technology, Reinventing Ihde*. Springer, pp. 141–155. doi: 10.1007/978-3-030-35967-6_9.

Winner, L. (2009) 'Do Artifacts Have Politics?', in Kaplan, D. M. (ed.) *Readings in the Philosophy of Technology*. 2nd edn. Rowman & Littlefield Publishers, Inc., pp. 251–263.

WIPO (2021) 'WIPO Technology Trends 2021: Assistive Technology'. Geneva: World Intellectual Property Organization. Available at: http://assistiveeducationaltechnology.weebly.com/assistive-technology.html#.

Žižek, S. (2011) *Violència*. Barcelona: Editorial Empúries.

Informe dels resultats

Aquesta secció ofereix una síntesi de les publicacions compilades, recollint-ne les idees principals i clarificant-ne el fil conductor.

En la primera d'elles, **"Prolegómenos a una ética para la robótica social"**, s'assenten les bases per a enfocar l'aproximació ètica a la robòtica social, a través d'una anàlisi filosòfica. En concret, s'endreça conceptualment la temàtica a examinar, amb desig d'evitar que la seva complexitat condueixi a desorientació del pensar.

D'entrada, davant l'ambigüitat terminològica generalitzada, s'ofereix una caracterització dels robots socials atenent al caràcter teleològicament subordinat de seva la interactivitat "social"; o sigui, al fet que la interacció és el *mitjà* a partir del qual realitzen una determinada tasca. A l'hora d'acotar-los com a objecte d'atenció ètica, aquest tret és clau per dos motius. D'una banda, perquè implica que les qüestions normativament rellevants de la interacció humà-robot no es poden pensar desvinculades de les finalitats a què respon la interacció. És a dir, no es pot problematitzar la interacció humà-robot en abstracte, desconsiderant el marc funcionalitat-finalitat a què aquesta respon i que, per norma general, s'associa a contextos pràctics específics. D'altra banda, perquè fa que sigui cabdal una reflexió hermenèutico-teleològica de les pràctiques relacionals a què serveixen els robots, per evitar dessubstancialitzar-les amb una excessiva reducció de la funció de la interacció social a l'acompliment de tasques.

Davant el desordre conceptual de la reflexió actual, es diluciden tres grans factors que augmenten la complexitat de la deliberació ètica sobre robòtica social. Primer, la multiplicitat d'esferes d'acció a què es pot orientar la reflexió, que tenen a veure no només amb l'entramat d'activitat humana implicada en el desplegament de la robòtica, sinó també amb l'agència (el "fer") dels robots d'IA. Segon, la condició instrumental de la tecnologia i, per tant, la subordinació teleològica del seu bé intern al d'altres activitats humanes, per la qual la reflexió ha procedir, en part, com a exercici d'ètica aplicada. Tercer, la convergència ontològica objecte-"subjecte" que es dona en els robots socials com a entitat de consideració ètica.

A més, s'identifica i analitza críticament el que es considera el nucli de la preocupació ètica entorn la robòtica social, designat com a "externalització de l'agència humana", i que té a veure amb la possibilitat de transferir, per a tasques en pràctiques relacionals, part de la nostra agència a aquests robots. En concret, s'adverteix que aquest fenomen, al focalitzar l'atenció en la dimensió agent dels robots i en qüestions primàriament relatives al seu "fer", pot generar reduccionismes en la mirada ètica. En aquest sentit, es revelen quatre característiques d'aquesta forma d'atansar-se a la reflexió normativa sobre la robòtica social:

(1) Una ontologia individualista dels robots, com a agents monàdics, que s'insereixen en la nostra quotidianitat en qualitat de subjectes o "quasi-altres"

(2) Un èmfasi de la reflexió ètica en el comportament d'aquests robots orientada a garantir la seva competència moral, o l'alineament del seu actuar amb els valors humans fonamentals i, per tant, configurada disciplinàriament com a "Machine ethics"

(3) La reducció del caràcter beneficiós d'un robot social a l'impacte del seu decidir i actuar en la tasca concreta que realitza

(4) Una limitació de l'atenció ètica a l'esfera individual de la vida humana, amb motiu del focus en l'impacte del "fer" del robot i, correlativament, en l'àmbit de la interacció diàdica humà-robot.

Contra el que seria aquesta aproximació ètica que es considera desenfocada, es defensa com a millor manera d' atansar-se a la robòtica social la centralitat de la pregunta pel "ser": abans de "fer correctament", un robot social ha de "ser (allò) correcte". Es tracta de la pregunta per la legitimitat. Es defineix així la "perspectiva del ser", i les seves categories de "teleologia" i "interès", com a fonament d'una aproximació ètica a la robòtica social més pertinent; més responsiva a l'agència subrogada dels robots i, per tant, més compromesa amb la pregunta per les finalitats, valors i interessos vinculats a la seva raó de ser.

S'esgrimeixen dos motius addicionals a favor d'aquesta perspectiva. En primer lloc, estructurar la mirada ètica des d'aquestes coordenades conceptuals és coherent amb el caràcter constitutivament instrumental de la robòtica social, que requereix pensar la funcionalitat dels robots (el seu "fer") des d'una atenció al marc teleològic i axiològic de la pràctica on s'insereixen. En segon lloc, adoptar la "perspectiva del ser" és coherent amb el caràcter d'una ètica que, compromesa amb la pregunta per la vida bona, reflexiona sobre la tecnologia en clau teleològica, antropològica i política, allunyant-se del que seria un mer exercici d' "impactologia" i d'una consideració normativa concernent merament al nivell *micro* de la vida humana.

Partint d'aquestes contribucions teòrico-conceptuals, en el segon article, **"The ethical issues of social robotics: A critical literature review"**, es facilita un diagnòstic de l'aproximació ètica a una de les branques centrals de la robòtica social, l'assistencial, a través d'una revisió crítica dels problemes ètics que se li associen en la discussió acadèmica contemporània.

D'aquesta revisió se n'extreuen tres dades importants sobre la reflexió ètica actual.

Primer, en la literatura sobre el tema s'identifiquen 26 problemes molt heterogenis i poc desenvolupats. En la majoria dels casos s'esmenten sense aclarir-los conceptualment o raonar-ne la seva "problematicitat". Malgrat aquesta heterogeneïtat, els problemes es poden classificar en tres grans grups temàtics –Benestar, Cura, Justícia–, segons l'esfera de la vida humana que es pren com a focus de les implicacions de la robòtica social assistencial –la individual, la de la pràctica on s'introdueixen els artefactes i la sociopolítica, respectivament–. Els problemes es poden classificar d'acord amb tres nivells de reflexió ètica: *micro-meso-macro*.

Segon, la major part (60%) dels problemes ètics tenen a veure amb les implicacions dels robots a nivell individual de la vida humana, en concret la de les persones que interactuen amb ells (els seus usuaris). Dins d'aquesta categoria s'hi troben els tres problemes que, per ser els més tractats, ocupen un lloc central en el panorama de discussió ètica actual: la privacitat, l'engany i l'autonomia (humana).

Tercer, el percentatge de problemes ètics relacionats amb les implicacions de la robòtica social assistencial per a les pràctiques en què s'insereixen els sistemes d'IA és considerablement baix (22%), així com també ho és, i de forma més accentuada, el percentatge de problemes sobre seves implicacions en l'estructuració sociopolítica de la vida humana (18%).

En base a aquests resultats, s'identifiquen certes característiques de la reflexió actual simptomàtiques d'una aproximació ètica deficient a la robòtica social assistencial.

En primer lloc, predomina una perspectiva ètica de tall individual, principalment centrada en les implicacions de la robòtica social assistencial a nivell *micro* per als usuaris d'aquests sistemes d'IA. Això va vinculat a una limitació del terreny de consideració critico-normativa a l'esfera de la interacció diàdica humà-robot. En canvi, romanen bastant desateses les implicacions de la tecnologia per a la pràctica on s'insereixen els artefactes (nivell *meso*), així com per a l'estructuració política de la societat (nivell *macro*). D'acord amb les consideracions teòrico-conceptuals ofertes i amb els fonaments per a l'ètica de la tecnologia explicats, la perspectiva ètica de focus individual és deficient perquè (1) implica una desatenció a la interrelació entre els nivells *micro-meso-macro* de la vida humana, un oblit de la dimensió política de la tecnologia i, en conseqüència, de la dimensió estructural de les relacions humà-robot; (2) entra en certa contradicció amb el caràcter constitutivament instrumental de la robòtica social i la necessitat d'una reflexió a la llum de la pràctica específica a què serveixen els artefactes; (3) converteix l'ètica en un exercici d'avaluació moral acrítica amb els fins i descompromesa de la pregunta pel tipus de vida i societats tecnològicament *mediades* que volem.

En segon lloc, hi ha una manca de reflexió sobre els pressupòsits teleològics i antropològics de la robòtica social assistencial, és a dir, sobre els significats relacionats amb la pràctica en què pretenen servir els robots ("assistència", "cura", "vulnerabilitat", "benestar", "autonomia", "capacitats). Aquesta manca de reflexió hermenèutica sobre les finalitats (el *per a què*) de la robòtica social assistencial suposa una deficiència important, principalment per dos motius: (1) el caràcter instrumental de la tecnologia demana d'una reflexió "aplicada" i, per tant, una hermenèutica crítica de les activitats humanes per a les quals es conceben els robots socials; (2) el caràcter mediador dels artefactes accentua la necessitat d'una reflexió crítica constant sobre les finalitats i els valors definitoris de l'activitat humana on els robots han de servir com a mitjans.

En tercer lloc, domina una comprensió restringida dels conceptes ètics nuclears al voltant dels quals s'articula la reflexió, com és el cas de les nocions de responsabilitat i llibertat. Això, que es correlaciona amb la perspectiva ètica individual predominant de l'aproximació actual, empobreix el terreny de qüestions normativament rellevants plantejades per la tecnologia.

En quart lloc, hi ha certes qüestions importants, en relació a les implicacions de la interacció humà-robot per al benestar individual de l'usuari, que romanen desateses. Aquestes concerneixen a la mediació tecnològica de la (inter)subjectivitat que suposa la relació d'alteritat humà-tecnologia pròpia de la robòtica social, i que es manté inexplorada per part de la filosofia de la tecnologia contemporània.

Finalment, a partir d'aquesta diagnosi es defineixen les línies generals d'una agenda de recerca de l'ètica en relació al camp de la robòtica social assistencial.

En el tercer article, **"Ethics for social robotics: A critical analysis"**, se sintetitzen i discuteixen aquelles tendències de la reflexió sobre robòtica social assistencial pròpies d'una aproximació ètica deficient a la tecnologia; a saber: (1) el predomini d'una perspectiva ètica individual, que oblida la dimensió política de la tecnologia; (2) la comprensió empobrida dels conceptes ètics en base als quals s'articula la reflexió, que redueix de forma dràstica el terreny de qüestions normativament rellevants per al desplegament tecnològic; i (3) la manca de reflexió sobre la teleologia de la robòtica social assistencial, que contravé el procedir de l'ètica aplicada que escau

a la tecnologia, atesa la seva condició instrumental i el seu caràcter "mediador".D'aquesta discussió se n'extreuen dues contribucions importants per al panorama de deliberació ètica actual sobre robòtica social.

Primer, es conclouen les bases d'una aproximació ètica apropiada a la robòtica social, que serà aquella que identifiqui i analitzi les implicacions d'aquesta tecnologia per als nivells *micro-meso-macro* de la vida humana. En termes de configuració disciplinària, això significa una ètica de la tecnologia ben articulada amb la filosofia política i la filosofia de la tecnologia.

Segon, en aquesta direcció i com a resposta als dèficits identificats, es concreten dues línies d'investigació per enriquir la reflexió ètica sobre la robòtica social per a contextos pràctics de cura, com és el cas de la branca assistencial: (1) estendre l'atenció ètica a les implicacions de la robòtica social per a la dimensió político-estructural de la llibertat; i (2) desplegar el potencial normatiu de la noció filosòfica de cura com a recurs per orientar la robòtica social assistencial.

En el quart article, **"Social assistive robotics: An ethical and political inquiry through the lens of freedom"**, es redefineixen les qüestions normativament rellevants per a un desplegament de la robòtica social assistencial compromès amb la llibertat humana. En particular, es desenvolupa la primera de les línies de recerca formulades en l'article anterior, és a dir, a través d'una anàlisi de les implicacions d'aquesta tecnologia des de la perspectiva filosòfica de llibertat.

Mitjançant un recorregut per diferents conceptualitzacions filosòfiques, es mostra que la llibertat, lluny de ser el mer exercici d'una facultat individual, és una capacitat humana constitutivament vinculada al marc sociopolític en què s'exerceix. En aquest sentit, es revela la dimensió político-estructural de la llibertat. Tanmateix, aquesta no figura com a objecte d'atenció ètica en la reflexió sobre les implicacions de la robòtica social assistencial per a la llibertat humana. Com s'evidencia amb una revisió de les problemàtiques ètiques en el marc de discussió acadèmica actual, aquestes s'articulen des d'una comprensió de llibertat en clau individual i s'entenen com a relatives a l'esfera de la interacció humà-robot –és a dir, com a problemes per a l'exercici de la llibertat dels individus amb qui interactuen els robots–.

Com a resultat d'una dilucidació del rol de la tecnologia en la configuració del marc sociopolític de la vida humana, es confirma la rellevància normativa de la dimensió político-estructural de la llibertat en relació al desplegament tecnològic. Per tant, s'evidencia la limitació de la reflexió ètica predominant al reduir el focus d'atenció a l'exercici individual de la llibertat, en desatenció al tipus de pràctiques, relacions de poder i estructures sociopolítiques que la robòtica social assistencial contribueix a vertebrar.

En aquest sentit, en resposta a aquesta limitada comprensió i com a primera contribució teòrico-pràctica, l'article avança la idea filosòfica de llibertat com a concepte ètic transversal per a la reflexió normativa sobre robòtica social assistencial.

A partir d'un reexamen de la robòtica social assistencial a la llum d'aquest concepte, s'identifiquen noves qüestions normativament rellevants per al seu desplegament que passen desapercebudes en el marc de discussió actual, i que tenen a veure amb el tipus de relacions intersubjectives, estructures de poder i capacitats humanes que la robòtica social assistencial fomenta. Com a segona contribució, l'article enriqueix el panorama de reflexió ètica actual,

ampliant el focus d'atenció més enllà de les implicacions de la interacció diàdica humà-robot i posant al centre la descuidada dimensió estructural de les relacions humà-robot.

Epíleg. Agenda per l'ètica de la robòtica social

Dels resultats de la investigació doctoral, es poden assenyalar un parell de línies per a l'agenda de l'ètica de la robòtica social, corresponents a dues tasques de l'ètica de les tecnologies: proporcionar coordenades normatives per al desplegament tecnocientífic i articular continguts per a la formació en ètica a professionals de l'enginyeria.

1. "Cura" i "Capacitats humanes" com a coordenades ètico-polítiques per a la robòtica social assistencial

El caràcter teleològicament subordinat i mediador de la tecnologia requereix, com s'ha indicat, d'una hermenèutica crítica de les activitats on han de servir els robots socials que permeti orientar-ne el seu desplegament adequadament. Atesos els dèficits que existeixen, en aquest sentit, en la reflexió ètica sobre la branca assistencial de la robòtica social (Pareto Boada, Román Maestre and Torras, 2022), urgeix incorporar la idea de "cura" en la recerca de l'ètica de la robòtica social. En particular, i per complementar les contribucions que ja s'han fet en aquesta línia amb el desenvolupament del Disseny Sensible als Valors Centrats en la Cura (CCVSD) (van Wynsberghe, 2013), convindria explorar el significat de cura des d'una perspectiva teleològica; és a dir, en termes de la seva finalitat (bé intern).

L'aportació diferencial d'una indagació filosòfica sobre el *per a què* de la cura –relatiu en última instància a la creació de "món" (Tronto, 1993, p.104)–, seria una ampliació del terreny de consideracions normativament rellevants per al desplegament tecnològic orientat a aquesta pràctica. En concret, suposaria una ampliació cap a qüestions relatives a la justícia i la vida bona, posant al centre la dimensió estructural de les relacions humà-robot. Això significaria (re)polititzar la cura com a pràctica relacional i com a concepte ètico-normatiu per al desplegament dels robots socials en l'àmbit assistencial, atès que la reflexió ètica predominant s'articula entorn una comprensió de cura com a relació intersubjectiva privada, com ho evidencien els problemes ètics del panorama de discussió actual (Pareto Boada, Román Maestre and Torras, 2021),

Al seu torn, la dimensió teleològica de la cura convida a explorar la idea filosòfica de "capacitats humanes" (Nussbaum, 2012) com a coordenada ètico-política per al disseny de robots socials per a l'assistència. Des d'una comprensió de l'assistència com a resposta a la vulnerabilitat humana no només en termes de facultats individuals sinó també de condicions estructurals, es transcendiria la comprensió restrictiva imperant de la tecnologia assistencial (Aparicio Payá *et al.*, 2019) com a substitutiva de funcionaments, per a concebre-la com a eina per crear la condició de possibilitat per a exercir-los (Toboso et al., 2020). En aquest sentit, de les diferents capacitats definides per Nussbaum (2012), n'hi ha tres que serien rectores específicament per a la mediació tecnològica de l'assistència mitjançant robots socials: Raó Pràctica, d'Afiliació i de Control sobre el propi entorn. Aquestes són essencials per a conduir la pròpia vida de forma autònoma, objectiu al qual pretén servir, en última instància, l'activitat assistencial.

2. Proposta docent en ètica per a l'enginyeria

La dimensió moral i política de la tecnologia, tal i com s'ha desplegat en la present tesi, té implicacions importants per a l'enginyeria i el tipus de formació ètica necessària per al seu

exercici professional. En particular, el caràcter mediador de la tecnologia situa l'enginyeria com a activitat inherentment moral, en tant que contribueix a donar resposta a la pregunta "què he de fer"? Lluny de tractar-se d'una mera producció de mitjans per a uns fins (externs) susceptibles de ser avaluats moralment, l'enginyeria és una activitat de disseny de mediacions tecnològiques i, per tant, de disseny de formes de vida. Conseqüentment, la responsabilitat ètica dels enginyers s'estén més enllà de les qüestions vinculades a la funcionalitat dels artefactes (usos i riscos), i engloba aquelles relatives al tipus de pràctiques i societats que la seva activitat contribueix a configurar.

En aquest sentit, es fa necessari dissenyar plans docents d'ètica per a graus universitaris en enginyeria que dotin als futurs professionals dels recursos adients per a poder assumir la responsabilitat ètica de la seva activitat. Es tracta d'ensenyar als enginyers a identificar quines són les qüestions normativament rellevants per al desplegament tecnològic des d'un punt de vista ètic, posicionar-s'hi críticament i respondre-hi en el seu exercici professional.

Sobretot en el context nacional català i espanyol, això suposa realitzar una tasca d'innovació docent, principalment per dos motius. En primer lloc, l'oferta de formació en ètica a les universitats politècniques s'hi troba encara en fase incipient, a diferència del que passa a altres països europeus i als Estats Units (Mitcham, 2009), on està consolidada de fa temps. En segon lloc, els pocs casos aïllats de sessions d'ètica que s'han donat fins ara en aquest tipus de graus han estat liderats i impartits, amb molt de mèrit, per perfils acadèmics d'enginyeria, al·lòctons a aquesta branca de la filosofia pràctica, seguint una tendència clarament internacional (Fiesler, Garrett and Beard, 2020). En un esforç interdisciplinari per articular els recursos critico-normatius de la filosofia amb el coneixement pràctic de cadascun dels camps tecnològics, es fa necessari implicar professionals de la filosofia i ètica de la tecnologia en la concepció i elaboració dels plans docents d'ètica per a enginyers.

En aquesta línia, s'adjunta una proposta de pla docent per a una assignatura de 6 crèdits ECTS d'ètica de la tecnologia per a l'enginyeria. Els seus continguts s'han començat a impartir a la Universitat Politècnica de Catalunya en el marc de l'assignatura optativa 'Ètica a la Ciència i l'Enginyeria-ECE' del 1r quadrimestre del curs 2023-24 de l'Escola Tècnica Superior d'Enginyeria de Telecomunicació de Barcelona.

En la proposta es presenten i descriuen breument els diferents blocs temàtics de l'assignatura, se'n detallen els seus continguts i les lectures corresponents. A més, s'introdueixen diferents casos de discussió i una llista de debats per desplegar la dimensió pràctica dels continguts teòrics tractats a classe i per reflexionar críticament sobre qüestions i problemes ètics vinculats a l'enginyeria. Aquesta llista s'acompanya de la bibliografia base de referència per a acotar el tema de debat.

Pla Docent - Ètica per a l'enginyeria

Descripció

En aquesta assignatura s'oferirà una introducció bàsica a conceptes i teories de l'ètica, a partir dels quals s'examinaran les implicacions de la ciència i la tecnologia per a les nostres vides, tant a nivell personal com per a la societat, les nostres relacions i les diferents activitats humanes. Les sessions seran sobretot pràctiques, i els coneixements teòrics es vincularan a l'anàlisi de problemes i controvèrsies actuals que els alumnes es poden trobar a la seva vida professional. Els objectius d'aquesta assignatura són (1) desenvolupar la capacitat de reflexionar sobre el significat ètic del desenvolupament tecnocientífic; i (2) poder sospesar i argumentar decisions complexes en la pròpia activitat professional.
No és necessari tenir coneixements previs en filosofia per a poder cursar aquesta assignatura.

Continguts

Tema 1. L'ètica: introducció
Aquest bloc temàtic ofereix una introducció bàsica a l'ètica com a disciplina filosòfica, sobretot en relació a la seva activitat, funcions, models teòrics, criteris i funcionament en la seva aplicació a camps d'activitat específics. Atesa la centralitat de l'argumentació en la reflexió ètica, es proporcionen coneixements elementals d'argumentació informal, relatius als criteris de bona argumentació, les fal·làcies en què es pot incórrer amb el seu incompliment i alguns principis de bones pràctiques per a la discussió ètica. L'objectiu d'aquesta primera familiarització amb l'ètica és doble: comprendre en què consisteix i què se'n pot esperar; i disposar d'unes bases per poder integrar-la en l'activitat professional.

1. Ètica: de què parlem?
2. Del "tramvia sense frens" al dilema del "cotxe autònom": un recorregut per les grans teories ètiques
◊Discussió 1**:** És moralment incorrecta la violència contra un gos robòtic?
3. L'ètica aplicada
4. Les quatre ètiques: individual, professional, organitzativa i cívica
5. L'argumentació crítica: elements bàsics
◊Discussió 2**:** Hauríem de fer campanya contra els robots sexuals (CASR)?

Lectures obligatòries

Vallor, S. (2015) 'Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character', *Philosophy and Technology*, 28(1), pp. 107–124. doi: 10.1007/s13347-014-0156-9. → *Ap. 1, 3 i 4*.

Richardson, K. (2015) 'The Asymmetrical "Relationship": Parallels Between Prostitution and the Development of Sex Robots'.

Lectures opcionals

Coeckelbergh, M. (2020) 'Should We Treat Teddy Bear 2.0 as a Kantian Dog? Four Arguments for the Indirect Moral Standing of Personal Social Robots, with Implications for Thinking About Animals and Humans', *Minds and Machines*. Springer Netherlands, (0123456789). doi: 10.1007/s11023-020-09554-3.

Tema 2. Ètica, per què? Sobre la dimensió moral i política de la tecnologia

Aquest bloc proporciona els coneixements per a respondre dues preguntes relacionades: per què la tecnologia demana d'ètica? i de quin tipus d'ètica es tracta? A partir d'un recorregut per contribucions centrals de la filosofia de la tecnologia clàssica i contemporània, s'exposa la dimensió moral i política de la tecnologia, especialment en la seva vinculació amb el caràcter mediador del artefactes. El propòsit d'aquest recorregut és descloure l'enginyeria com a activitat moral i definir corresponentment la responsabilitat ètica dels enginyers, que té a veure amb el seu rol proactiu en el disseny de formes de vida, de relacions de poder i d'estructura sociopolítica.

1. Tecnologia i ètica: una relació amb història
  1.1. El principi de responsabilitat
  1.2. La qüestió dels riscos
  ◊Discussió 3: Una moratòria a la IA?
2. Tecnologia: aproximacions de concepte
3. La dimensió moral de la tecnologia
  3.1. La tecnologia, un mer instrument?
  3.2. La teoria de la mediació tecnològica
    3.2.1. Les relacions humà-tecnologia i la doble mediació tecnològica
    3.2.1. La mediació tecnològica de la moralitat
    3.2.3. De la teoria de la mediació a l'enginyeria com a activitat ètica
    ◊Discussió 4: "Moralitzar la tecnologia"? La proposta d' H. Achterhuis
  3.3. Canvi Tecno-moral: com succeeix?
4. La dimensió política de la tecnologia

Lectures obligatòries

Jonas, H. (2015) *El principio de responsabilidad. Ensayo de una ética para la civilización tecnológica.* Herder. Barcelona. → *Pròleg*.

Innerarity, D. (2023) 'Una moratoria artificial', *El País*, 24 April. Available at: https://elpais.com/opinion/2023-04-24/una-moratoria-artificial.html .

Floridi, L. (2023) 'On Good and Evil , the Mistaken Idea That Technology Is Ever Neutral , and the Importance of the Double - Charge Thesis', *Philosophy & Technology*. Springer Netherlands, pp. 1–5. doi: 10.1007/s13347-023-00661-4. → *Introducció*.

Verbeek, P. P. (2008) 'Obstetric ultrasound and the technological mediation of morality: A postphenomenological analysis', *Human Studies*, 31(1), pp. 11–26. doi: 10.1007/s10746-007-9079-0.

Danaher, J. and Sætra, H. S. (2023) 'Mechanisms of Techno-Moral Change: A Taxonomy and Overview', *Ethical Theory and Moral Practice*. doi: 10.1007/s10677-023-10397-x. → *Secció 3*.

Winner, L. (2009) 'Do Artifacts Have Politics?', in Kaplan, D. M. (ed.) *Readings in the Philosophy of Technology*. 2nd edn. Rowman & Littlefield Publishers, Inc., pp. 251–263. → *Introducció*.

Lectures opcionals

Esquirol i Calaf, J. M. (2011) 'TÉCNICA Y HUMANISMO: CUATRO MIRADAS FILOSÓFICAS', *Argumentos de Razón Técnica*, 14, pp. 69–86. → *Ap. Hans Jonas*.

Ortega y Gasset, J. (2004) '¿Qué es la técnica?', en *Meditación de la técnica y otros ensayos sobre ciencia y filosofía*. 8ª. Revista de Occidente en Alianza Editorial.

Esquirol i Calaf, J. M. (2006) 'Heidegger i la tècnica', *Cercles: revista d'història cultural*, (9), pp. 83–101.

Verbeek, P.-P. (2015) 'Beyond Interaction: A Short Introduction to Mediation Theory', *Interactions*, 22(3), pp. 26–31. doi: 10.1145/2751314.

Tema 3. Conceptes ètics fonamentals

Aquest bloc avança conceptes centrals per la reflexió ètica dels enginyers, per a poder identificar i avaluar críticament les qüestions normativament rellevants del desplegament tecnològic, sobretot en vistes a la presa de decisions en l'exercici professional. Es tracta d'un bloc de continguts amb finalitats pràctiques, que proporciona als enginyers recursos per a entendre el significat del que està en joc amb la seva activitat, i afrontar i lidiar amb la seva responsabilitat ètica. S'ofereixen conceptes ètics transversals, que permeten pensar en les implicacions de la tecnologia per als diferents nivells *micro-meso-macro* de la vida humana, i que convé reexaminar perquè alguns d'ells podria desdibuixar-se amb la introducció de tecnologies d'IA.

1. Responsabilitat
- 1.1. Clarificacions conceptuals
- 1.2. Els "(Techno)-Responsibility gaps"
  - 1.1.1. El "problema de les múltiples mans" i la responsabilitat col·lectiva
  - 1.1.2. Tecnologies d'IA autònomes i "Control Humà Significatiu" (MHC)
  - ◊Discussió 5: Els buits de tecno-responsabilitat moral, un problema o una solució?
- 1.2.3. Responsabilitat narrativa

2. Llibertat
- 2.1. Dos conceptes de llibertat
- 2.2. L'autonomia (humana)
- 2.3. Llibertat com a no-dominació

3. Capacitats humanes: L'enfoc d'A. Sen i M. Nussbaum

4. Justícia
- 4.1. Justícia distributiva: distribuir què, com i a qui?
- 4.2. Discriminació algorítmica i injustícia estructural
- 4.3. Algoritmes i injustícia epistèmica

5. Estudi de cas: la robòtica social assistencial

◊Discussió 6: La IA i l'aproximació ètica per riscos de la Unió Europea

Lectures obligatòries

Danaher, J. (2023) 'The Case for Outsourcing Morality to AI', *WIRED*, pp. 1–10. Available at: https://www.wired.com/story/philosophy-artificial-intelligence-responsibility-gap/.

Domingo Moratalla, T. (2013) 'Què significa "autonomia moral"? Hannah Arendt', *Bioètica & Debat*.

Berlin, I. (2017) 'Dos conceptos de libertad', en *Sobre la libertad*. Alianza Editorial.

Nussbaum, M. C. (2012) 'Las capacidades centrales', en *Crear capacidades. Propuesta para el desarrollo humano*. Paidós.

van Wynsberghe, A. (2021) 'Sustainable AI: AI for sustainability and the sustainability of AI', *AI and Ethics*. Springer International Publishing, 1(3), pp. 213–218. doi: 10.1007/s43681-021-00043-6.

Young, I. M. (2019) 'Political Responsibility and Structural Injustice', in E. Goodin, R. and Pettit, P. (eds) *Contemporary Political Philosophy: An Anthology*. 3rd Editio. Wiley-Blackwell, pp. 253–262. → *Introducció.*

Lectures opcionals

Coeckelbergh, M. (2021) 'Máquinas *arresponsables* y decisiones inexplicables', en *Ética de la inteligencia artificial*. Cátedra.

Coeckelbergh, M. (2021) 'Narrative responsibility and artificial intelligence', *AI & SOCIETY*. Springer London, (0123456789). doi: 10.1007/s00146-021-01375-x.

Pettit, P. (2020) 'Antes de la libertad negativa y la libertad positiva', en *Republicanismo. Una teoría sobre la libertad y el gobierno*. 9ª. Paidós.

Coeckelbergh, M. (2023) 'Igualdad y justícia: sesgo y discriminación por la IA', en *La filosofía política de la inteligencia artificial. Una introducción*. Cátedra.

Tema 4. Metodologies d'integració de l'ètica en la pràctica tecnocientífica

Aquest bloc temàtic proporciona alguns recursos i metodologies per a integrar l'ètica en l'activitat de l'enginyeria, desglossant-los en relació a tres dimensions o àmbits de la pràctica tecnocientífica: institucional, disseny tecnològic i recerca. Els objectius principals d'aquesta secció són dos: facilitar mètodes per al disseny deliberat i responsable de mediacions tecnològiques; i proporcionar les bases per a una ètica de l'enginyeria que no es redueixi a una responsabilitat professional entesa només en clau individual, sinó que es desplegui com a ètica organitzacional o institucional.

1. La infraestructura institucional: Codis ètics, Codis de conducta i Comitès d'ètica
2. Ètica del disseny: El "Disseny Sensible als Valors" (VSD)
  2.1. El Disseny Sensible als Valors de Cura
  2.2. El Disseny Sensible a les Capacitats
3. Ètica de la recerca

Lectures obligatòries

Jacobs, N. (2020) 'Capability Sensitive Design for Health and Wellbeing Technologies', *Science and Engineering Ethics*. Springer Netherlands, 26(6), pp. 3363–3391. doi: 10.1007/s11948-020-00275-5.

Lectures opcionals

Verbeek, P.-P. (2011) 'Morality in Design', in *Moralizing Technology: Understanding and Designing the Morality of Things*. The University of Chicago Press.

Tronto, J. (1993) 'Chapter 4: Care', in *Moral Boundaries. A Political Argument for an Ethic of Care*. Routledge.

van Wynsberghe, A. (2013) 'Designing Robots for Care: Care Centered Value-Sensitive Design', *Science and Engineering Ethics*, 19(2), pp. 407–433. doi: 10.1007/s11948-011-9343-6.

Proposta de debats

1. Són els "griefbots" una tecnologia per al dol moralment acceptable?

2. *Nudging* tecnològic per a millorar-nos moralment: una forma admissible de paternalisme?

3. Robots d'IA sexuals: hauríem de dissenyar-los amb capacitat de consentiment sexual?

4. Tecnologies d'IA socialment interactives i "emocionals": una erosió de la dignitat humana?

5. És moralment lícit dissenyar els fills "a la carta", a través de l'enginyeria genètica?

6. Pot ser la meritocràcia un bon model de justícia algorítmica?

Bibliografia de referència [Debats]

Jiménez-Alonso, B. and Brescó de Luna, I. (2023) 'Griefbots. A New Way of Communicating With The Dead?', *Integrative Psychological and Behavioral Science*, 57(2), pp. 466–481. doi: 10.1007/s12124-022-09679-3.

Borenstein, J. and Arkin, R. (2016) 'Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being', *Science and Engineering Ethics*. Springer Netherlands, 22(1), pp. 31–46. doi: 10.1007/s11948-015-9636-2.

Frank, L. and Nyholm, S. (2017) 'Robot sex and consent: Is consent to sex between a robot and a human conceivable, possible, and desirable?', *Artificial Intelligence and Law*. Springer Netherlands, 25(3), pp. 305–323. doi: 10.1007/s10506-017-9212-y.

Gutiu, S. (2012) 'Sex Robots and Roboticization of Consent', in *We Robot Conference 2012*.

Savulescu, J. (2002) 'Deaf lesbians, "designer disability," and the future of medicine', *British Medical Journal*, 325(7367), pp. 771–773. doi: 10.1136/bmj.325.7367.771.

Sandel, M. (2015) *Contra la perfección. La ética en la era de la ingeniería genética*. 2ª. Barcelona: Marbot Ediciones.

Binns, R. (2017) 'Fairness in Machine Learning: Lessons from Political Philosophy', (2016), pp. 1–11. Available at: http://arxiv.org/abs/1712.03586.

Sandel, M. (2020) *La tiranía del mérito: ¿Qué ha sido del bien común?* Debate.

Referències

Aparicio Payá, M. *et al.* (2019) 'Un marco ético-político para la robótica asistencial. An Ethical-Political Framework for Assistive Robotics', *ArtefaCTos. Revista de estudios de la ciencia y la tecnología*, 8(1), pp. 97–117.

Fiesler, C., Garrett, N. and Beard, N. (2020) 'What Do We Teach When We Teach Tech Ethics?', pp. 289–295. doi: 10.1145/3328778.3366825.

Mitcham, C. (2009) 'A historico-ethical perspective on engineering education: From use and convenience to policy engagement', *Engineering Studies*, 1(1), pp. 35–53. doi: 10.1080/19378620902725166.

Nussbaum, M. C. (2012) *Crear capacidades. Propuesta para el desarrollo humano*. Paidós.

Pareto Boada, J., Román Maestre, B. and Torras, C. (2021) 'The ethical issues of social assistive robotics: A critical literature review', *Technology in Society*, 67. doi: 10.1016/j.techsoc.2021.101726.

Pareto Boada, J., Román Maestre, B. and Torras, C. (2022) 'Ethics for social robotics: A critical analysis', in *TRAITS Workshop Proceedings (arXiv:2206.08270) held in conjunction with Companion of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*. Springer Berlin Heidelberg, pp. 1284–1286.

Toboso, M. *et al.* (2020) 'Robotics as an Instrument for Social Mediation', *Biosystems and Biorobotics*, 25, pp. 51–58. doi: 10.1007/978-3-030-24074-5_11.

Tronto, J. (1993) *Moral Boundaries. A Political Argument for an Ethic of Care*. Routledge.

van Wynsberghe, A. (2013) 'Designing Robots for Care: Care Centered Value-Sensitive Design', *Science and Engineering Ethics*, 19(2), pp. 407–433. doi: 10.1007/s11948-011-9343-6.

CONCLUSIONS

Aquesta investigació doctoral dona resposta a la necessitat urgent d'orientar, des de l'ètica, el desplegament de la robòtica social, en especial pel que fa a la seva branca assistencial. Aquesta necessitat ve donada, d'una banda, per l'emergència d'aquest camp de la robòtica intel·ligent com a proveïdora d'eines per l'àmbit de la salut, i, de l'altra, per les insuficiències de la reflexió ètica que per ara l'acompanya, i que fonamentalment tenen a veure amb una mala articulació disciplinària de l'ètica en relació a la robòtica social. Això s'explica pel fet que la reflexió ha estat liderada sobretot per perfils al·lòctons a aquesta branca de la filosofia pràctica, però també per una vaguetat generalitzada sobre l'estatut disciplinari de l'ètica de la tecnologia i, per tant, sobre el seu procedir, que correspon a la pròpia ètica clarificar.

La present tesi s'ha estructurat a l'entorn de dos grans objectius.

El primer, establir un marc conceptual per a una correcta aproximació ètica a la robòtica social, és a dir, per a identificar i analitzar les qüestions normativament rellevants per al seu desenvolupament, sobretot per a l'àmbit assistencial. En aquesta línia, s'han assentat els fonaments disciplinaris per a l'abordatge ètic d'aquesta tecnologia (Cap. 2), a través d'una clarificació de tres qüestions clau: (1) per què la tecnologia demana d'ètica?; (2) de quin tipus d'ètica es tracta?; i (3) quin és l'estatut de l'ètica de la tecnologia?

1) Per què la tecnologia demana d'ètica?

Una bona aproximació ètica a la tecnologia exigeix comprendre el sentit en què la tecnologia entra en joc amb l'àmbit de la moralitat i concerneix a l'activitat de fonamentació critico-racional d'aquesta. En aquest sentit, es diluciden dues dimensions, moral i política, de la tecnologia que la situen com a èticament significativa.

Contra el que pressuposa la concepció instrumental pròpia de la Modernitat –clarament bastant vigent encara en el sector de l'enginyeria contemporània–, la tecnologia no és moralment neutral. Per la seva condició instrumental constitutivament vinculada a un substrat hermenèutico-pràctic concret, i pel caràcter *mediador* dels artefactes en les relacions humà-mon, la tecnologia no es pot seguir pensant des d'una òptica ontològica moderna que la separa de l'àmbit de la moralitat. Al contrari: com recentment assenyala la postfenomenologia (Verbeek, 2011), en tant que, en el seu ús, els artefactes contribueixen a configurar les nostres percepcions i interpretacions de la realitat, i les nostres accions i pràctiques, la tecnologia participa activament en la moralitat. I ho fa tant a nivell *micro* com *macro*, doncs també configura els marcs interpretatius i morals de la societat.

La tecnologia té també una dimensió política, en el sentit que juga un paper decisiu en l'organització sociopolítica de la vida humana, en la disposició relacional dels individus en el si de la comunitat. Això s'explica no només per la funcionalitat instrumental dels artefactes, en virtut de la qual les tecnologies poden ser usades amb fins polítics concrets. Més fonamentalment, la dimensió política de les tecnologies es deu al seu caràcter *mediador*: al co-determinar els marcs interpretatius i les pràctiques humanes, els artefactes organitzen relacions de poder tant a nivell intersubjectiu *micro* com a nivell social, *macro*.

De la indagació sobre aquesta primera qüestió se'n deriven tres conclusions importants per a l'ètica de la robòtica social:

a) No procedeix entendre la tecnologia, en la seva qualitat d'activitat i d'artefacte, com a concernent a l'ètica només amb motiu de les finalitats i riscos que s'hi vinculen, sinó també amb motiu del tipus de pràctiques i societats que contribueix a configurar.

b) Lluny de ser una mera activitat de producció de mitjans per a uns fins (externs), l'enginyeria és una activitat inherentment moral, de disseny de mediacions tecnològiques (Verbeek, 2006) i, per tant, de modes de vida.

c) Les relacions humà-tecnologia no s'han d'entendre només en termes d'ús o d'interacció directa amb els artefactes, sinó també en la seva dimensió estructural: ens relacionem amb les tecnologies no només en tant que usuaris, sinó també de forma més indirecta com a subjectes en estructures tecnològicament *mediades*.

2) De quin tipus d'ètica es tracta?

No es tracta d'una ètica externa a la tecnologia, ocupada de determinar l'acceptabilitat moral dels seus usos i riscos. Degut a la dimensió moral i política de la tecnologia, ha de ser una ètica integrada en el seu desenvolupament des de la fase inicial de concepció i disseny tecnològic. Per aquest mateix motiu, és una ètica positiva que, més enllà de demarcar límits, delibera sobre el tipus de subjectivitats i societats a configurar.

3) Quin és l'estatut de l'ètica de la tecnologia?

Es defineix l'ètica de la tecnologia com una ètica *parcialment* aplicada *subsidiària* de les activitats humanes a què es destinen els artefactes. Per la condició instrumental de la tecnologia, cal contextualitzar la reflexió en el marc de finalitats i valors de les activitats per a les quals les tecnologies es conceben com a mitjà, per això es tracta d'una ètica aplicada subsidiària. Ara bé, aquest caràcter "aplicat" de l'ètica de la tecnologia és només parcial. Més enllà d'una reflexió a la llum de les pràctiques on serveixen, la dimensió moral i política de la tecnologia requereix aproximar-la des d'una perspectiva crítica, d'acord amb la qual es delibera no només sobre mitjans sinó també sobre els fins. L'ètica de la tecnologia no pot ser merament una ètica pragmàtica prudencial (Rorty, 1997), centrada en l'eficiència per aconseguir els fins ja donats o preestablerts a la pràctica, sinó una ètica filosòfica que, d'acord amb l'afany d'emancipació que la defineix, reflexioni sobre la tecnologia en clau teleològica, antropològica i política.

D'aquesta definició se'n deriven dues conclusions importants per a la configuració disciplinària de l'ètica de les tecnologies i, entre elles, de la robòtica social:

d) Contra la proliferació contemporània d'ètiques regionals per als diferents camps tecnològics (Sætra and Danaher, 2022), no procedeix crear una ètica per a cada tecnologia. Això no s'escau amb el tipus d'aplicació subsidiària que demana la tecnologia, d'acord amb la qual la demarcació de dominis d'estudi no pot respondre tant a les tecnologies com a les activitats humanes a què aquestes pretenen servir.

e) L'ètica de la tecnologia s'ha d'articular com una triangulació disciplinar entre ètica aplicada, filosofia política i filosofia de la tecnologia.

Amb aquesta primera tasca de fonamentació disciplinària per a l'abordatge ètic de la tecnologia, la tesi contribueix a articular l'ètica de la robòtica social com a camp de l'ètica contemporània, superant la vaguetat generalitzada sobre el seu estatut i procedir. Alhora, s'invalida l'actual comprensió restrictiva de l'ètica de l'enginyeria com una mera ètica de la professió (Franssen, Lokhorst and van de Poel, 2023), desvinculada de la reflexió sobre el tipus de pràctiques i societats que es configuren amb la mediació tecnològica. També es proporcionen les bases teòriques per a superar els dèficits que presenta la manera com s'ha estat enfocant la reflexió fins ara.

Aquests dèficits, identificats a partir d'una revisió crítica dels problemes ètics associats a la robòtica social assistencial (Cap.3), són propis d'una ètica de la tecnologia disfuncional, que deixa fora del terreny de consideració normativa la dimensió política de les tecnologies i que, a nivell metodològic, no procedeix en conformitat amb l'estatut d'ètica aplicada que li pertocaria. En particular, en el panorama de discussió actual predomina una perspectiva ètica individual, centrada en les implicacions de la robòtica per als usuaris d'aquests sistemes d'IA i, per tant, limitada a l'esfera de la interacció diàdica humà-robot. La deliberació critico-normativa s'està fent en clau *micro*, deixant bastant desateses les implicacions de la robòtica social assistencial per a les pràctiques en què s'insereixen els robots (nivell *meso*) i per a l'estructuració de la societat en el seu conjunt (nivell *macro*). A més, no es discuteixen obertament i de forma crítica els pressupòsits teleològics i antropològics de la robòtica social assistencial.

En vistes al primer dels objectius de tesi, els fonaments disciplinaris per l'ètica de la tecnologia s'han complementat amb la delimitació d'unes coordenades ètiques específiques per a la robòtica social assistencial (Cap. 3, 4); a saber: llibertat, cura i capacitats humanes. D'una banda, es tracta de tres conceptes filosòfics centrals per a aproximar aquesta tecnologia de forma contextualitzada, identificant i analitzant les qüestions normativament rellevants per al seu desplegament a la llum de les finalitats i valors vinculats a l'assistència –com així ho requereix l'estatut d'ètica aplicada subsidiària que li és propi a l'ètica de la tecnologia–. D'altra banda, aquests tres conceptes són transversals, porten a pensar en les implicacions de la robòtica social assistencial en tots els nivells *micro-meso-macro* de la vida humana, que estan interrelacionats. En aquest sentit, són coordenades des de les quals adreçar pertinentment la dimensió política de la tecnologia.

El segon objectiu de la tesi ha estat (re)analitzar qüestions ètiques centrals de la robòtica social assistencial des del marc conceptual establert (Cap. 3). En aquesta línia, l'abast ètico-normatiu de la idea filosòfica de llibertat ha permès redefinir les qüestions normativament rellevants per al desplegament de la robòtica social assistencial, ampliant el focus d'atenció cap a la dimensió política de la tecnologia. En concret, s'afegeixen al terreny de consideració normativa qüestions relatives al tipus de tipus de relacions intersubjectives, estructures de poder i capacitats humanes que la robòtica social assistencial fomenta.

La tesi contribueix així a desplegar la robòtica social assistencial des del compromís amb la llibertat humana, la qual cosa demana estendre el focus d'atenció més enllà de les implicacions

de la interacció diàdica humà-robot, posant al centre la descuidada dimensió estructural de les relacions humà-robot.

Les altres dues coordenades ètico-polítiques per a la robòtica social assistencial, cura i capacitats humanes, es dibuixen com a futures línies per l'agenda de l'ètica de la robòtica social (Cap. 3: Epíleg). Tenint en compte l'estat de la literatura a partir del marc teòric de l'ètica de la cura de J. Tronto i de l'enfoc de les capacitats de M. Nussbaum, s'avancen unes indicacions per a desplegar la força normativa d'aquests conceptes filosòfics.

Finalment, contra una comprensió limitada de l'ètica de l'enginyeria com a ètica de la professió, i davant la necessitat d'una innovació docent en ètica per graus universitaris d'enginyeria que doti als futurs professionals de recursos adequats per a fer-se càrrec de la dimensió moral i política de la seva activitat, la tesi proposa un pla docent per a una assignatura d'ètica de la tecnologia de 6 crèdits ECTS (Cap.3: Epíleg).

Referències

Franssen, M., Lokhorst, G.-J. and van de Poel, I. (2023) 'Philosophy of Technology', *The Stanford Encyclopedia of Philosophy (Spring 2023 Edition)*. Edward N. Zalta & Uri Nodelman (eds.). Available at: https://plato.stanford.edu/archives/spr2023/entries/technology/.

Rorty, R. (1997) *¿ESPERANZA O CONOCIMIENTO? Una introducción al pragmatismo*. Fondo de Cultura Económica.

Sætra, H. S. and Danaher, J. (2022) 'To Each Technology Its Own Ethics : The Problem of Ethical Proliferation', *Philosophy & Technology*. Springer Netherlands, 35(93), pp. 1–26. doi: 10.1007/s13347-022-00591-7.

Verbeek, P.-P. (2006) 'Materializing Morality. Design Ethics and Technological Mediation', *Science, Technology, & Human Values*, 31(3), pp. 361–380.

Verbeek, P.-P. (2011) *Moralizing Technology: Understanding and Designing the Morality of Things*. The University of Chicago Press.

BIBLIOGRAFIA

Ajuntament de Barcelona (2020) Misty II the social robot becomes part of the lives of twenty senior citizens. Available at: https://www.barcelona.cat/infobarcelona/en/tema/senior-citizens/misty-ii-the-social-robot-becomes-part-of-the-lives-of-twenty-senior-citizens_907645.html (Accessed: 31 July 2021).

Allen, C., Wallach, W. and Smit, I. (2006) 'Why machine ethics?', IEEE Intelligent Systems, 21(4), pp. 12–17. doi: 10.1109/MIS.2006.83.

Andriella, A., Torras, C. and Alenyà, G. (2019) 'Short-Term Human–Robot Interaction Adaptability in Real-World Environments', International Journal of Social Robotics. Springer Netherlands. doi: 10.1007/s12369-019-00606-y.

Andriella, A., Torras, C. and Alenyà, G. (2020) 'Cognitive System Framework for Brain-Training Exercise Based on Human-Robot Interaction', Cognitive Computation. doi: 10.1007/s12559-019-09696-2.

Aparicio Payá, M. et al. (2019) 'Un marco ético-político para la robótica asistencial. An Ethical-Political Framework for Assistive Robotics', ArtefaCTos. Revista de estudios de la ciencia y la tecnología, 8(1), pp. 97–117.

Aparicio, M. et al. (2020) 'Discursive Frameworks for the Development of Inclusive Robotics', Biosystems and Biorobotics, 25, pp. 74–80. doi: 10.1007/978-3-030-24074-5_14.

Arnold, T. and Scheutz, M. (2017) 'Beyond Moral Dilemmas: Exploring the Ethical Landscape in HRI', ACM/IEEE International Conference on Human-Robot Interaction, Part F1271, pp. 445–452. doi: 10.1145/2909824.3020255.

Arras, J. D. (1990) 'Review: Common Law Morality', The Hastings Center Report, 20(4), pp. 35–37.

Battistuzzi, L. et al. (2020) 'Socially Assistive Robots, Older Adults and Research Ethics: The Case for Case-Based Ethics Training', International Journal of Social Robotics. Springer Netherlands. doi: 10.1007/s12369-020-00652-x.

Baumer, E. P. S. and Silberman, M. S. (2011) 'When the implication is not to design (technology)', Conference on Human Factors in Computing Systems - Proceedings, pp. 2271–2274. doi: 10.1145/1978942.1979275.

Bayertz, K. (2003) 'La moral como construcción. Una autorreflexión sobre la ética aplicada', in Cortina, A. and García Marzá, D. (eds) Razón Pública y éticas aplicadas: los caminos de la razón práctica en una sociedad pluralista. Tecnos, pp. 47–70.

Bekey, G. A. (2014) 'Current Trends in Robotics: Technology and Ethics', in Lin, P., Abney, K., and Bekey, G. A. (eds) Robot Ethics. The Ethical and Social Implications of Robotics. MIT Press, pp. 17–34.

Berlin, I. (1969) 'Two Concepts of Liberty', in Four Essays on Liberty. Oxford University Press.

Berlin, I. (1969) Four Essays on Liberty. Oxford University Press.

Bisconti Lucidi, P. and Nardi, D. (2018) 'Companion Robots: The Hallucinatory Danger of Human-Robot Interactions', AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 17–22. doi: 10.1145/3278721.3278741.

Borenstein, J. and Arkin, R. C. (2017) 'Nudging for good: robots and the ethical appropriateness of nurturing empathy and charitable behavior', AI and Society. Springer London, 32(4), pp. 499–507. doi: 10.1007/s00146-016-0684-1.

Breazeal, C., Takanishi, A. and Kobayashi, T. (2008) 'Social Robots that Interact with People', in Siciliano, B. and Khatib, O. (eds) Springer Handbook of Robotics. Springer, pp. 1349–1369. doi: https://doi.org/10.1007/978-3-540-30301-5_59.

Brey, P. (2010) 'Philosophy of Technology after the Empirical Turn', Techné, 14(1), pp. 36–48.

Brundage, M. (2014) 'Limitations and risks of machine ethics', Journal of Experimental and Theoretical Artificial Intelligence, 26(3), pp. 355–372. doi: 10.1080/0952813X.2014.895108.

Busquets Surribas, M. (2019) 'Descubriendo la importancia ética del cuidado', Folia humanística, (12).

Caleb-Solly, P. (2016) 'A brief introduction to … Assistive robotics for independent living', Perspectives in Public Health, 136(2), pp. 70–72.

Caleb-Solly, P. et al. (2014) 'A mixed-method approach to evoke creative and holistic thinking about robots in a home environment', ACM/IEEE International Conference on Human-Robot Interaction, pp. 374–381. doi: 10.1145/2559636.2559681.

Campa, R. and Campa, R. (2016) 'The rise of social robots : a review of the recent literature', Journal of Evolution and Technology, 26(1).

Camps, V. (2017) Breve historia de la ética. Barcelona: RBA Libros.

Cappuccio, M. L., Peeters, A. and McDonald, W. (2020) 'Sympathy for Dolores: Moral Consideration for Robots Based on Virtue and Recognition', Philosophy and Technology. Philosophy & Technology, 33(1), pp. 9–31. doi: 10.1007/s13347-019-0341-y.

Cave, S. et al. (2019) 'Motivations and Risks of Machine Ethics', Proceedings of the IEEE. Institute of Electrical and Electronics Engineers Inc., 107(3), pp. 562–574. doi: 10.1109/JPROC.2018.2865996.

Chita-Tegmark, M. and Scheutz, M. (2021) 'Assistive Robots for the Social Management of Health: A Framework for Robot Design and Human–Robot Interaction Research', International Journal of Social Robotics. Springer Netherlands, 13(2), pp. 197–217. doi: 10.1007/s12369-020-00634-z.

Coeckelbergh, M. (2009) 'Personal Robots, Appearance, and Human Good: A Methodological Reflection on Roboethics', International Journal of Social Robotics, 1(3), pp. 217–221. doi: 10.1007/s12369-009-0026-2.

Coeckelbergh, M. (2011) 'Is ethics of robotics about robots? Philosophy of robotics beyond realism and individualism', Law, Innovation and Technology, 3(2), pp. 241–250. doi: 10.5235/175799611798204950.

Coeckelbergh, M. (2011) 'You, robot: On the linguistic construction of artificial others', AI and Society, 26(1), pp. 61–69. doi: 10.1007/s00146-010-0289-z.

Coeckelbergh, M. (2012) 'Are emotional robots deceptive?', IEEE Transactions on Affective Computing. IEEE, 3(4), pp. 388–393. doi: 10.1109/T-AFFC.2011.29.

Coeckelbergh, M. (2015) 'Artificial agents, good care, and modernity', Theoretical Medicine and Bioethics. Kluwer Academic Publishers, 36(4), pp. 265–277. doi: 10.1007/s11017-015-9331-y.

Coeckelbergh, M. (2018) 'Technology and the good society: A polemical essay on social ontology, political principles, and responsibility for technology', Technology in Society. Elsevier Ltd, 52, pp. 4–9. doi: 10.1016/j.techsoc.2016.12.002.

Coeckelbergh, M. (2020) 'Should We Treat Teddy Bear 2.0 as a Kantian Dog? Four Arguments for the Indirect Moral Standing of Personal Social Robots , with Implications for Thinking About Animals and Humans', Minds and Machines. Springer Netherlands, (0123456789). doi: 10.1007/s11023-020-09554-3.

Coeckelbergh, M. (2020) AI Ethics. MIT Press.

Coeckelbergh, M. (2020) Introduction to Philosophy of Technology. Oxford University Press.

Coeckelbergh, M. (2021) 'Narrative responsibility and artificial intelligence', AI & SOCIETY. Springer London, (0123456789). doi: 10.1007/s00146-021-01375-x.

Coeckelbergh, M. (2022) Robot Ethics. The MIT Press.

Coeckelbergh, M. (2022) The Political Philosophy of AI . Polity Press.

Coeckelbergh, M. et al. (2016) 'A Survey of Expectations About the Role of Robots in Robot-Assisted Therapy for Children with ASD: Ethical Acceptability, Trust, Sociability, Appearance, and Attachment', Science and Engineering Ethics. Springer Netherlands, 22(1), pp. 47–65. doi: 10.1007/s11948-015-9649-x.

Coeckelbergh, M. et al. (eds) (2018) 'Envisioning Robots in Society - Power, Politics, and Public Space', in Proceedings of Robophilosophy 2018 /TRANSOR 2018. IOS Press. doi: 10.1017/CBO9781107415324.004.

Comisión Europea (2018) Comunicación de la Comisión al Parlamento Europeo, al Consejo Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones. Plan coordinado sobre inteligencia artificial.

Corporació Catalana de Mitjans Audiovisuals (2023) 'ARI, el robot que facilita tasques a la gent gran i arribarà a un miler de llars catalanes', 24 November. Available at: https://www.ccma.cat/324/ari-el-robot-que-facilita-tasques-a-la-gent-gran-i-arribara-a-un-miler-de-llars-catalanes/noticia/3262661/.

Cortina, A. (1996) 'El estatuto de la ética aplicada. Hermenéutica crítica de las actividades humanas', Isegoría, 13, pp. 119–134.

Cortina, A. (2003) 'El quehacer público de las éticas aplicadas: ética cívica transnacional', in Cortina, A. and García-Marzá, D. (eds) Razón pública y éticas aplicadas. Los caminos de la razón práctica en una sociedad pluralista. Tecnos, pp. 13–44.

Cortina, A. (2007) Ética mínima. Introducción a la filosofía práctica. Tecnos. Madrid.

Cortina, A. and Martínez, E. (2001) Ética. 3a Ed. Akal.

Craglia, M. et al. (2018) Artificial Intelligence - A European perspective. Luxembourg: Joint Research Center. doi: 10.2760/11251.

Crawford, K. (2021) Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press.

Damiano, L. and Dumouchel, P. (2018) 'Anthropomorphism in Human-Robot Co-evolution', Frontiers in Psychology, 9(MAR), pp. 1–9. doi: 10.3389/fpsyg.2018.00468.

de Graaf, M. M. A. (2016) 'An Ethical Evaluation of Human–Robot Relationships', International Journal of Social Robotics. Springer Netherlands, 8(4), pp. 589–598. doi: 10.1007/s12369-016-0368-5.

de Pagter, J. (2023) 'From EU Robotics and AI Governance to HRI Research: Implementing the Ethics Narrative', International Journal of Social Robotics. doi: 10.1007/s12369-023-00982-6.

Didier, C. and Heriard-Dubreuil, B. (2005) 'Engineering Ethics in Europe', in Mitcham, C. (ed.) Encyclopedia of Science, Technology and Ethics. Macmillan Reference USA, pp. 632–635.

Dignum, V. et al. (2018) 'Design for Values for Social Robot Architectures', Frontiers in Artificial Intelligence and Applications, 311(January 2019), pp. 43–52. doi: 10.3233/978-1-61499-931-7-43.

Dolic, Z., Castro, R. and Moarcas, R. (2019) Robots in healthcare: a solution or a problem?, Study for the Committee on Environment, Public Health, and Food Safety, European Parliament.

Duffy, M. (2007) 'Doing the dirty work: Gender, race, and reproductive labor in historical perspective', Gender and Society, 21(3), pp. 313–336. doi: 10.1177/0891243207300764.

Echeverría, J. (2003) La revolución tecnocientífica. Madrid: Fondo de Cultura Económica de España.

Esquirol, J. M. (2011) Los filósofos contemporáneos y la técnica. De Ortega a Sloterdijk. Barcelona: Editorial Gedisa.

European Commission (2020) White Paper on Artificial Intelligence - A European approach to excellence and trust. doi: 10.1017/CBO9781107415324.004.

European Commission (2021) 'Proposal for a Regulation of the European Parliament and of the Council. Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts'.

European Parliament (2017) Civil Law Rules on Robotics. European Parliament resolution of 16 February 2017 with recomendations to the Commission on Civil Law Rules on Robotics (2015/2013(INL)), Official Journal of the European Union. Available at: http://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html#title10.

Feenberg, A. (2009a) 'Democratic Rationalization: Technology, Power, and Freedom', in Kaplan, D. M. (ed.) Readings in the Philosophy of Technology. 2nd ed. Rowman & Littlefield Publishers, Inc.

Feenberg, A. (2009b) 'Peter-Paul Verbeek: Review of What Things Do', Human Studies, 32(2), pp. 225–228. doi: 10.1007/s10746-009-9115-3.

Feenberg, A. (2010) Between Reason and Experience: Essays in Technology and Modernity. Cambridge: MIT Press.

Feenberg, A. (2018) 'What Is Philosophy of Technology?', in Beira, E. and Feenberg, A. (eds) Tecnology, Modernity, and Democracy. Essays by Andrew Feenberg. Rowman & Littlefield International.

Feil-Seifer, D. and Matarić, M. J. (2005) 'Defining Socially Assistive Robotics', in 9th International Conference on Rehabilitation Robotics. IEEE, pp. 465–468.

Feil-Seifer, D. and Matarić, M. J. (2011) 'Socially Assistive Robotics: Ethical Issues Related to Technology', IEEE Robotics and Automation Magazine, 18(1), pp. 24–31. doi: 10.1109/MRA.2010.940150.

Fernández-Aller, C. et al. (2021) 'An Inclusive and Sustainable Artificial Intelligence Strategy for Europe Based on Human Rights', IEEE Technology and Society Magazine, (March). doi: 10.1109/MTS.2021.3056283.

Fiesler, C., Garrett, N. and Beard, N. (2020) 'What Do We Teach When We Teach Tech Ethics?', pp. 289–295. doi: 10.1145/3328778.3366825.

Fiske, A., Henningsen, P. and Buyx, A. (2019) 'Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy', Journal of Medical Internet Research, 21(5), pp. 1–12. doi: 10.2196/13216.

Floridi, L. and Cowls, J. (2019) 'A Unified Framework of Five Principles for AI in Society', Harvard Data Science Review, (1), pp. 1–13. doi: 10.1162/99608f92.8cd550d1.

Fong, T., Nourbakhsh, I. and Dautenhahn, K. (2003) 'A survey of socially interactive robots', Robotics and Autonomous Systems, 42(3–4), pp. 143–166. doi: 10.1016/S0921-8890(02)00372-X.

Fosch-Villaronga, E. and Grau Ruiz. María Amparo (2019) 'Expert Considerations for the Regulation of Assistive Robotics. A European Robotics Forum Echo', Dilemata, Revista Internacional de Éticas Aplicadas, (30), pp. 149–169.

Foucault, M. (1977) Discipline and Punish: The Birth of the Prison. New York: Vintage Books.

Franssen, M., Lokhorst, G.-J. and van de Poel, I. (2023) 'Philosophy of Technology', The Stanford Encyclopedia of Philosophy (Spring 2023 Edition). Edward N. Zalta & Uri Nodelman (eds.). Available at: https://plato.stanford.edu/archives/spr2023/entries/technology/.

Fundació iSocial (2020) Misty II, robot per millorar la qualitat de vida de persones grans que viuen soles.

Fundació Víctor Grífols i Lucas (2019) 'Un robot permet que el cuidador tingui més temps per fer tasques amb valor emocional'. Available at: https://www.fundaciogrifols.org/ca/-/entrevista-a-carme-torras (Accessed: 20 November 2023).

Funk, M. and Coeckelbergh, M. (2020) (Technical) Autonomy as Concept in Robot Ethics, Biosystems and Biorobotics. Springer. doi: 10.1007/978-3-030-24074-5_12.

General Assembly of the United Nations (1948) Universal Declaration of Human Rights.

Generalitat de Catalunya (2023) El robot que dona de menjar a pacients al programa de Els Matins de TV3. Available at: https://perevirgili.gencat.cat/ca/detalls/Noticia/El-robot-que-dona-de-menjar-a-pacients-al-programa-de-Els-Matins-de-TV3 (Accessed: 30 November 2023).

González García, M. I. and Fernández-Jimeno, N. (2022) 'Introducción. La filosofía de la tecnología y sus identidades múltiples. Una mirada desde España', Azafea. Revista de Filosofía. Monográfico. Cuestiones actuales en Filosofía de la Tecnología, 24, pp. 7–19.

Goodrich, M. A. and Schultz, A. C. (2007) 'Human-Robot interaction: A Survey', Foundations and Trends in Human-Computer Interaction, 1(3), pp. 203–275. doi: 10.1561/1100000005.

Gordon, J.-S. and Nyholm, S. (2023) 'Ethics of Artificial Intelligence', The Internet Encyclopedia of Philosophy. Available at: https://iep.utm.edu/ethics-of-artificial-intelligence/.

Grupo de expertos de alto nivel sobre inteligencia artificial (2019) Directrices éticas para una IA fiable. doi: 10.2759/14078.

Gunkel, D. J. (2018) 'The other question: can and should robots have rights?', Ethics and Information Technology. Springer Netherlands, 20, pp. 87–99. doi: 10.1007/s10676-017-9442-4.

Haring, K. S. et al. (2019) 'The Dark Side of Human-Robot Interaction: Ethical Considerations and Community Guidelines for the Field of HRI', 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, pp. 689–690. Available at: http://arxiv.org/abs/1504.04339.

Hasan, R. (2021) 'Republicanism and Structural Domination', Pacific Philosophical Quarterly, 102(2), pp. 292–319. doi: 10.1111/papq.12337.

Heidegger, M. (1954) 'La pregunta por la técnica', Revista de filosofía, pp. 34–41.

Heidegger, M. (2009) 'The Question Concerning Technology', in Kaplan, D. M. (ed.) Readings in the Philosophy of Technology. 2nd edn. Rowman & Littlefield Publishers, Inc., pp. 9–24.

Heidegger, M. (2021) La pregunta por la técnica. 1a. Herder.

Heuer, T., Schiering, I. and Gerndt, R. (2018) 'Privacy and Socially Assistive Robots - A Meta Study', in Privacy and Identity Management. The Smart Revolution. Springer International Publishing, pp. 265–281. doi: 10.1007/978-3-319-92925-5.

High-Level Expert Group on AI (2019) 'Ethics Guidelines for Trustworthy AI'. European Commission, pp. 1–41.

High-Level Expert Group on Artificial Intelligence (2019) 'A definition of AI: Main capabilities and scientific disciplines. Definition developed for the purpose of the AI HLEG's deliverables'. European Commission, p. 7. Available at: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56341.

Huber, A., Weiss, A. and Rauhala, M. (2016) 'The Ethical Risk of Attachment: How to Identify, Investigate and Predict Potential Ethical Risks in the Development of Social Companion Robots', ACM/IEEE International Conference on Human-Robot Interaction, 2016-April, pp. 367–374. doi: 10.1109/HRI.2016.7451774.

Hui, Y. (2020) Fragmentar el futuro: ensayos sobre tecnodiversidad. Buenos Aires: Caja Negra.

Ienca, M. et al. (2016) 'Social and Assistive Robotics in Dementia Care: Ethical Recommendations for Research and Practice', International Journal of Social Robotics. Springer Netherlands, 8(4), pp. 565–573. doi: 10.1007/s12369-016-0366-7.

Ihde, D. (1990) Technology and the Lifeworld. Indiana University Press.

Ihde, D. (2009) 'A Phenomenology of Technics', in Kaplan, D. M. (ed.) Readings in the Philosophy of Technology. 2nd Ed. Rowman & Littlefield Publishers, Inc.

Ihde, D. (2015) Postfenomenología y Tecnociencia.Conferencias en la Universidad de Pekín. Plataforma Editorial Sello.

Illa Mestre, M. (2018) Proposta d'una polisèmia estructurada del concepte <<Intimitat>>. Universitat de Barcelona.

Jackson, R. B. and Williams, T. (2019) 'Language-Capable Robots may Inadvertently Weaken Human Moral Norms', ACM/IEEE International Conference on Human-Robot Interaction. IEEE, 2019-March, pp. 401–410. doi: 10.1109/HRI.2019.8673123.

Johnson G., D. (2011) 'Computer Systems: Moral Entities but Not Moral Agents', in Anderson, M. and Anderson, S. L. (eds) Machine Ethics. Cambridge University Press, pp. 168–183. doi: https://doi.org/10.1017/CBO9780511978036.

Jonas, H. (2015) El principio de responsabilidad. Ensayo de una ética para la civilización tecnológica. Herder. Barcelona.

Kaplan, D. M. (2009) 'What Things Still Don't Do', Human Studies, 32(2), pp. 229–240. doi: 10.1007/s10746-009-9116-2.

Kapp, E. (1877) Grundlinien einer Philosophie der Technik. Zur Entstehungsgeschichte der Kultur aus neuen Gesichtspunkten. Reprint. 2015 Hamburg: Felix Meiner Verlag.

Kettner, M. (2003) 'Tres Dilemas Estructurales de la Ética Aplicada', in Cortina, A. and García-Marzá, D. (eds) Razón pública y éticas aplicadas. Los caminos de la razón práctica en una sociedad pluralista. Tecnos, pp. 145–158.

Koimizu, J. (2019) 'Aged Care with Socially Assistive Robotics under Advance Care Planning', Proceedings of IEEE Workshop on Advanced Robotics and its Social Impacts, ARSO. IEEE, 2019-Octob, pp. 34–38. doi: 10.1109/ARSO46408.2019.8948742.

Körtner, T. (2016) 'Ethical challenges in the use of social service robots for elderly people', Zeitschrift fur Gerontologie und Geriatrie, 49(4), pp. 303–307. doi: 10.1007/s00391-016-1066-5.

Koyama, T. (2016) 'Ethical Issues for Social Robots and the Trust-based Approach', Proceedings of IEEE Workshop on Advanced Robotics and its Social Impacts, ARSO. IEEE, 2016-Novem, pp. 1–5. doi: 10.1109/ARSO.2016.7736246.

Kudina, O. (2019) The technological mediation of morality: value dynamism, and the complex interaction between ethics and technology. PhD Thesis, University of Twente.

Kudina, O. and Verbeek, P. P. (2019) 'Ethics from Within: Google Glass, the Collingridge Dilemma, and the Mediated Value of Privacy', Science Technology and Human Values, 44(2), pp. 291–314. doi: 10.1177/0162243918793711.

Latour, B. (1993) We Have Never Been Modern. Cambridge: Harvard University Press.

Latour, B. (1994) 'On Technical Mediation: Philosophy, Sociology, Genealogy', Common Knowledge, Fall V3(2), pp. 29–64.

Lehoux, P. and Grimard, D. (2018) 'When robots care: Public deliberations on how technology and humans may support independent living for older adults', Social Science and Medicine. Elsevier, 211(June), pp. 330–337. doi: 10.1016/j.socscimed.2018.06.038.

Liedo, B. (2021) 'Vulnerabilidad', Eunomía. Revista en Cultura de la Legalidad, 20, pp. 242–257. doi: https://doi.org/10.20318/eunomia.2021.6074.

Liedo, B. and Ausín Díez, T. (2022) 'Alcance y límites de la tecnologización del cuidado: aprendizajes de una pandemia', Revista Espanola de Salud Pública, 96.

Lin, P. (2014) 'Introduction to Robot Ethics', in Lin, P., Abney, K., and Bekey, G. A. (eds) Robot Ethics. The Ethical and Social Implications of Robotics. The MIT Press, pp. 3–15.

López Aranguren, J. L. (1991) De ética y de moral. Barcelona: Círculo de Lectores.

López Aranguren, J. L. (1994) Ética. Obras completas, II. Madrid: Trotta.

López Aranguren, J. L. (2005) Ética. Madrid: Alianza.

Maalouf, N. et al. (2018) 'Robotics in Nursing: A Scoping Review', Journal of Nursing Scholarship, 50(6), pp. 590–600. doi: 10.1111/jnu.12424.

MacIntyre, A. (1984) 'Does applied ethics rest on a mistake?', The Monist, 67(4), pp. 498–513. doi: http://www.jstor.org/stable/27902885.

MacIntyre, A. (2001) Animales racionales y dependientes. Por qué los seres humanos necesitamos las virtudes. Barcelona: Paidós.

MacIntyre, A. (2007) After Virtue. A Study in Moral Theory. 3rd ed. University of Notre Dame Press.

MacIntyre, A. (2019) Tras la virtud. Barcelona: Austral.

Mackenzie, C. (2021) 'Relational Autonomy', in Hall, K. Q. and Ásta (eds) The Oxford Handbook of Feminist Philosophy. Oxford Academic. doi: https://doi.org/10.1093/oxfordhb/9780190628925.013.29.

Mackenzie, C. and Stoljar, N. (eds) (2000) Relational Autonomy: Feminist Perspectives on Autonomy, Agency and the Social Self. Oxford University Press.

Marzano, M. (2009) Consiento, luego existo. Ética de la autonomía, Proteus. Proteus.

Matarić, M. J. (2017) 'Socially assistive robotics: Human augmentation versus automation', Science Robotics, 2(4). doi: 10.1126/scirobotics.aam5410.

Matarić, M. J. and Scassellati, B. (2016) 'Socially Assistive Robotics', in Siciliano, B. and Khatib, O. (eds) Springer Handbook of Robotics. Springer, pp. 1973–1994.

McBride, N. (2020) 'Robot Enhanced Therapy for Autistic Children: An Ethical Analysis', IEEE Technology and Society Magazine. IEEE, 39(1), pp. 51–60. doi: 10.1109/MTS.2020.2967493.

Mejia, C. and Kajikawa, Y. (2017) 'Bibliometric Analysis of Social Robotics Research: Identifying Research Trends and Knowledgebase', Applied Sciences (Switzerland), 7(12). doi: 10.3390/app7121316.

Millar, J. (2015) 'Technology as Moral Proxy: Autonomy and Paternalism by Design', IEEE Technology and Society Magazine, 34(2), pp. 47–55. doi: 10.1109/MTS.2015.2425612.

Miller, L. F. (2020) 'Human Rights of Users of Humanlike Care Automata', Human Rights Review. Human Rights Review, 21(2), pp. 181–205. doi: 10.1007/s12142-020-00581-2.

Misselhorn, C., Pompe, U. and Stapleton, M. (2013) 'Ethical Considerations Regarding the Use of Social Robots in the Fourth Age', GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry, 26(2), pp. 121–133. doi: 10.1024/1662-9647/a000088.

Mitcham, C. (1994) Thinking Through Technology. The Path between Engineering and Philosophy. The University of Chicago Press.

Mitcham, C. (2009) 'A historico-ethical perspective on engineering education: From use and convenience to policy engagement', Engineering Studies, 1(1), pp. 35–53. doi: 10.1080/19378620902725166.

Mitcham, C. and Briggle, A. (2009) The Interaction of Ethics and Technology in Historical Perspective, Philosophy of Technology and Engineering Sciences. Elsevier B.V. doi: 10.1016/B978-0-444-51667-1.50045-8.

Morozov, E. (2013) To save everything, click here. The folly of technological solutionism. Public Affairs.

Noori, F. M., Uddin, Z. and Torresen, J. (2019) 'Robot-Care for the Older People: Ethically Justified or Not?', 2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob). IEEE, pp. 43–47. doi: 10.1109/DEVLRN.2019.8850706.

Nordmann, A. and Rip, A. (2009) 'Mind the gap revisited', Nature Nanotechnology. Nature Publishing Group, 4, pp. 273–274. doi: 10.1038/nnano.2009.26.

Nørskov, M., Seibt, J. and Santiago Quick, O. (eds) (2020) 'Culturally Sustainable Social Robotics', in Proceedings of Robophilosophy 2020. IOS Press.

Nussbaum, M. C. (2012) Crear capacidades. Propuesta para el desarrollo humano. Paidós.

Nylander, S., Ljungblad, S. and Jimenez Villareal, J. (2012) 'A complementing approach for identifying ethical issues in care robotics - Grounding ethics in practical use', Proceedings - IEEE International Workshop on Robot and Human Interactive Communication. IEEE, pp. 797–802. doi: 10.1109/ROMAN.2012.6343849.

O'Brolcháin, F. (2019) 'Robots and people with dementia: Unintended consequences and moral hazard', Nursing Ethics, 26(4), pp. 962–972. doi: 10.1177/0969733017742960.

Operto, F. and Veruggio, G. (2008) 'Roboethics: Social and Ethical Implications of Robotics', in Siciliano, B. and Khatib, O. (eds) Springer Handbook of Robotics. Springer, pp. 1499–1524. doi: 10.1007/978-3-540-30301-5.

Ortega y Gasset, J. (2004) Meditación de la técnica y otros ensayos sobre ciencia y filosofía. 8a. Revista de Occidente en Alianza Editorial.

Page, M. J. et al. (2021) 'The PRISMA 2020 statement: An updated guideline for reporting systematic reviews', The BMJ, 372. doi: 10.1136/bmj.n71.

Pareto Boada, J. (2021) 'Prolegómenos a una ética para la robótica social', Dilemata, Revista Internacional de Éticas Aplicadas, (34), pp. 71–87.

Pareto Boada, J., Román Maestre, B. and Torras, C. (2021) 'The ethical issues of social assistive robotics: A critical literature review', Technology in Society, 67. doi: 10.1016/j.techsoc.2021.101726.

Pareto Boada, J., Román Maestre, B. and Torras, C. (2022) 'Ethics for social robotics: A critical analysis', in TRAITS Workshop Proceedings (arXiv:2206.08270) held in conjunction with Companion of the 2022 ACM/IEEE International Conference on Human-Robot Interaction. Springer Berlin Heidelberg, pp. 1284–1286.

Parlamento Europeo (2017) 'Normas de Derecho Civil sobre robótica. Resolución del Parlamento Europeo, de 16 de febrero de 2017, con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica (2015/2013(INL))'. Diario Oficial de la Unión Europea.

Payr, S. M. (2015) 'Towards Human-Robot Interaction Ethics', in Trappl, R. (ed.) A Construction Manual for Robots' Ethical Systems. Cognitive Technologies. Springer. doi: 10.1007/978-3-319-21548-8.

Pettit, P. (2002) Republicanism. A Theory of Freedom and Government. Oxford University Press.

Pitt, J. (2014) 'Guns Don't Kill, People Kill; Values in and/or around Technologies', in The Moral Status of Technical Artifacts. Springer, pp. 89–101.

Plató (1988) Diàlegs, vol. IX. Barcelona: Fundació Bernat Metge.

Pranckutė, R. (2021) 'Web of science (Wos) and Scopus: The Titans of Bibliographic Information in Today's Academic World', Publications, 9(12), pp. 1–59. doi: 10.3390/publications9010012.

Rabbitt, S. M., Kazdin, A. E. and Scassellati, B. (2015) 'Integrating socially assistive robotics into mental healthcare interventions: Applications and recommendations for expanded use', Clinical Psychology Review. Elsevier B.V., 35, pp. 35–46. doi: 10.1016/j.cpr.2014.07.001.

Rawls, J. (2005) Political Liberalism. Expanded E. New York: Columbia University Press.

Richardson, K. (2019) 'The human relationship in the ethics of robotics: a call to Martin Buber's I and Thou', AI and Society. Springer London, 34(1), pp. 75–82. doi: 10.1007/s00146-017-0699-2.

Ricoeur, P. (2008) Lo justo 2. Estudios, lecturas y ejercicios de ética aplicada. Madrid: Trotta.

Robeyns, I. (2017) 'Clarifications', in Wellbeing, Freedom and Social Justice: The Capability Approach Re-Examined. Open Books, pp. 89–168.

Robson, A. (2018) 'Intelligent machines, care work and the nature of practical reasoning', Nursing Ethics, 26, pp. 1906–1916. doi: 10.1177/0969733018806348.

Rodogno, R. (2016) 'Ethics and social robotics', Ethics and Information Technology. Springer Netherlands, 18(4), pp. 241–242. doi: 10.1007/s10676-016-9412-2.

Rodogno, R. (2016) 'Social robots, fiction, and sentimentality', Ethics and Information Technology. Springer Netherlands, 18(4), pp. 257–268. doi: 10.1007/s10676-015-9371-z.

Román Maestre, B. (2016) Ética de los servicios sociales. Herder.

Román Maestre, B. (2021) 'Llibertat. Idees clàssiques pel món que ve. 1/3'. Available at: https://www.instituthumanitats.org/ca/documents/videos/1-3-idees-classiques-per-al-mon-que-ve-llibertat-sessio-1.

Rorty, R. (1997) ¿ESPERANZA O CONOCIMIENTO? Una introducción al pragmatismo. Fondo de Cultura Económica.

Rosenberger, R. (2017) Callous Objects: Designs Against the Homeless. University of Minnesota Press.

Rosenberger, R. and Verbeek, P.-P. (2015) 'A Field Guide to Postphenomenology', in Rosenberger, R. and Verbeek, P.-P. (eds) Postphenomenological Investigations: Essays on Human-Technology Relations. Lexington Books.

Rosenberger, R. and Verbeek, P.-P. (eds) (2015) Postphenomenological Investigations: Essays on Human–Technology Relations. Lexington Books.

Ruiz Trujillo, P. (2020) Ética de las nanotecnologías. Herder.

Sætra, H. S. and Danaher, J. (2022) 'To Each Technology Its Own Ethics : The Problem of Ethical Proliferation', Philosophy & Technology. Springer Netherlands, 35(93), pp. 1–26. doi: 10.1007/s13347-022-00591-7.

Sandel, M. (2015) Contra la perfección. La ética en la era de la ingeniería genética. 2a. Barcelona: Marbot Ediciones.

Santoni de Sio, F. and van Wynsberghe, A. (2016) 'When Should We Use Care Robots? The Nature-of-Activities Approach', Science and Engineering Ethics. Springer Netherlands, 22(6), pp. 1745–1760. doi: 10.1007/s11948-015-9715-4.

Sarrica, M., Brondi, S. and Fortunati, L. (2020) 'How many facets does a "social robot" have? A review of scientific and popular definitions online', Information Technology and People, 33(1), pp. 1–21. doi: 10.1108/ITP-04-2018-0203.

Schoeman, F. D. (ed.) (1984) Philosophical Dimensions of Privacy: An Anthology, Philosophical Dimensions of Privacy.

Seibt, J. (2017) 'Robophilosophy', in Braidotti, R. and Hlavajova, M. (eds) Posthuman Glossary. London: Bloomsbury Academic, pp. 390–393.

Seibt, J., Nørskov, M. and Schack Andersen, S. (eds) (2016) 'What Social Robots Can and Should Do', in Proceedings of Robophilosophy 2016 / TRANSOR 2016. IOS Press.

Sharkey, A. (2014) 'Robots and human dignity: A consideration of the effects of robot care on the dignity of older people', Ethics and Information Technology, 16(1), pp. 63–75. doi: 10.1007/s10676-014-9338-5.

Sharkey, A. (2020) 'Can we program or train robots to be good?', Ethics and Information Technology. Springer, 22, pp. 283–295. doi: 10.1007/s10676-017-9425-5.

Sharkey, A. and Sharkey, N. (2012) 'Granny and the robots: Ethical issues in robot care for the elderly', Ethics and Information Technology, 14(1), pp. 27–40. doi: 10.1007/s10676-010-9234-6.

Sheridan, T. B. (2020) 'A review of recent research in social robotics', Current Opinion in Psychology. Elsevier Ltd, 36, pp. 7–12. doi: 10.1016/j.copsyc.2020.01.003.

Sorell, T. and Draper, H. (2017) 'Second thoughts about privacy, safety and deception', Connection Science, 29(3), pp. 217–222. doi: 10.1080/09540091.2017.1318826.

Sparrow, R. (2016) 'Robots in aged care: a dystopian future?', AI and Society. Springer London, 31(4), pp. 445–454. doi: 10.1007/s00146-015-0625-4.

Sparrow, R. (2019) 'Robotics Has a Race Problem', Science, Technology, & Human Values, p. 016224391986286. doi: 10.1177/0162243919862862.

Sparrow, R. and Sparrow, L. (2006) 'In the hands of machines? The future of aged care', Minds and Machines, 16(2), pp. 141–161. doi: 10.1007/s11023-006-9030-6.

Stahl, B. C. and Coeckelbergh, M. (2016) 'Ethics of healthcare robotics: Towards responsible research and innovation', Robotics and Autonomous Systems. Elsevier B.V., 86, pp. 152–161. doi: 10.1016/j.robot.2016.08.018.

Stokes, F. and Palmer, A. (2020) 'Artificial Intelligence and Robotics in Nursing: Ethics of Caring as a Guide to Dividing Tasks Between AI and Humans', Nursing Philosophy, (May), pp. 1–9. doi: 10.1111/nup.12306.

Tapus, A., Matarić, M. and Scassellati, B. (2007) 'The Grand Challenges in Socially Assistive Robotics', IEEE Robotics and Automation Magazine, 14(1), pp. 35–42.

Thaler, R. H. and Sunstein, C. R. (2009) Nudge: Improving Decisions about Health, Wealth, and Happiness. London: Penguin.

The Member States (2012) Charter of Fundamental Rights of the European Union, Official Journal of the European Union. doi: 10.2307/j.ctt1ffjmjq.33.

Toboso, M. et al. (2020) 'Robotics as an Instrument for Social Mediation', Biosystems and Biorobotics, 25, pp. 51–58. doi: 10.1007/978-3-030-24074-5_11.

Torralba, F. (2009) La intimitat. Pagès Editors.

Torras, C. (2019) 'Assistive Robotics : Research Challenges and Ethics Education Initiatives', Dilemata, Revista Internacional de Éticas Aplicadas, (30), pp. 63–77.

Torras, C. (2019) 'Social networks and robot companions: Technology, ethics, and science fiction', Metode. Universitat de Valencia, 2019(9), pp. 163–169. doi: 10.7203/metode.9.12479.

Torras, C. (2024) 'Ethics of Social Robotics: Individual and Societal Concerns and Opportunities', Annual Review of Control, Robotics, and Autonomous Systems, 7(1), pp. 1–18. doi: 10.1146/annurev-control-062023-082238.

Torras, C. and Ludescher, L. G. (2023) 'Writing Science Fiction as an Inspiration for AI Research and Ethics Dissemination', Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 13500 LNAI, pp. 322–344. doi: 10.1007/978-3-031-24349-3_17.

Tronto, J. (1993) Moral Boundaries. A Political Argument for an Ethic of Care. Routledge.

Tzafestas, S. G. (2016) 'Socialized Roboethics', in Roboethics. A Navigating Overview. Springer.

Tzafestas, S. G. (2016) Roboethics. A Navigating Overview. Springer.

Tzafestas, S. G. (2018) 'Roboethics: Fundamental concepts and future prospects', Information (Switzerland), 9(6). doi: 10.3390/INFO9060148.

UNESCO (2019) I'd blush if I could: Closing gender divides in digital skills through education. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1.

United Nations (2019) World population prospects 2019. Highlights, Department of Economic and Social Affairs, Population Division. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12283219.

United Nations Department of Economic and Social Affairs (2022) World Population Prospects 2022. Summary of Results. Available at: www.un.org/development/ desa/pd/.

Vallès-Peris, N. (2021) 'Repensar la robótica y la inteligencia artificial desde la ética de los cuidados', Teknokultura. Revista de Cultura Digital y Movimientos Sociales, 18(2), pp. 137–146. doi: 10.5209/tekn.73983.

Vallès-Peris, N. and Domènech, M. (2020) 'Roboticists' Imaginaries of Robots for Care: The Radical Imaginary as a Tool for an Ethical Discussion', Engineering Studies, 12(3), pp. 157–176. doi: 10.1080/19378629.2020.1821695.

Vallès-Peris, N. and Domènech, M. (2020b) 'ROBOTS PARA LOS CUIDADOS. LA ÉTICA DE LA ACCIÓN MESURADA FRENTE A LA INCERTIDUMBRE.', Cuadernos de bioetica : revista oficial de la Asociacion Espanola de Bioetica y Etica Medica, 31(101), pp. 87–100. doi: 10.30444/CB.54.

Vallès-Peris, N., Angulo, C. and Domènech, M. (2018) 'Children's Imaginaries of Human-Robot Interaction in Healthcare', International Journal of Environmental Research and Public Health. MDPI AG, 15(5). doi: 10.3390/ijerph15050970.

Vallor, S. (2015) 'Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character', Philosophy and Technology, 28(1), pp. 107–124. doi: 10.1007/s13347-014-0156-9.

Vallverdú, J. and Casacuberta, D. (2015) 'Ethical and technical aspects of emotions to create empathy in medical machines', in van Rysewyk, S. P. and Pontier, M. (eds) Machine Medical Ethics. Springer International Publishing, pp. 341–362. doi: 10.1007/978-3-319-08108-3_20.

Van Aerschot, L. and Parviainen, J. (2020) 'Robots responding to care needs? A multitasking care robot pursued for 25 years, available products offer simple entertainment and instrumental assistance', Ethics and Information Technology. Springer Netherlands, (0123456789). doi: 10.1007/s10676-020-09536-0.

van der Plas, A., Smits, M. and Wehrmann, C. (2010) 'Beyond speculative robot ethics: A vision assessment study on the future of the robotic caretaker', Accountability in Research, 17(6), pp. 299–315. doi: 10.1080/08989621.2010.524078.

van Maris, A. et al. (2020) 'Designing Ethical Social Robots—A Longitudinal Field Study With Older Adults', Frontiers in Robotics and AI, 7(January). doi: 10.3389/frobt.2020.00001.

Van Maris, A. et al. (2020) 'The Impact of Affective Verbal Expressions in Social Robots', ACM/IEEE International Conference on Human-Robot Interaction, pp. 508–510. doi: 10.1145/3371382.3378358.

van Wynsberghe, A. (2013) 'Designing Robots for Care: Care Centered Value-Sensitive Design', Science and Engineering Ethics, 19(2), pp. 407–433. doi: 10.1007/s11948-011-9343-6.

van Wynsberghe, A. (2016) 'Service robots, care ethics, and design', Ethics and Information Technology. Springer Netherlands, 18(4), pp. 311–321. doi: 10.1007/s10676-016-9409-x.

van Wynsberghe, A. and Li, S. (2019) 'A paradigm shift for robot ethics: from HRI to human–robot–system interaction (HRSI)', Medicolegal and Bioethics, 9, pp. 11–21. doi: 10.2147/mb.s160348.

Vandemeulebroucke, T., Casterle, B. D. and Gastmans, C. (2020) 'Ethics of socially assistive robots in aged-care settings: A socio-historical contextualisation', Journal of Medical Ethics, 46(2), pp. 128–136. doi: 10.1136/medethics-2019-105615.

Vandemeulebroucke, T., Dierckx de Casterlé, B. and Gastmans, C. (2018) 'The use of care robots in aged care: A systematic review of argument-based ethics literature', Archives of Gerontology and Geriatrics. Elsevier, 74(August 2017), pp. 15–25. doi: 10.1016/j.archger.2017.08.014.

Vanderelst, D. and Winfield, A. (2018) 'The Dark Side of Ethical Robots', AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, (December), pp. 317–322. doi: 10.1145/3278721.3278726.

Véliz, C. (2020) Privacy is Power: Why and How You Should Take Back Control of Your Data. Bantam Press.

Verbeek, P. P. (2008) 'Obstetric ultrasound and the technological mediation of morality: A postphenomenological analysis', Human Studies, 31(1), pp. 11–26. doi: 10.1007/s10746-007-9079-0.

Verbeek, P. P. (2020) 'Politicizing Postphenomenology', in Miller, G. and Shew, A. (eds) Reimagining Philosophy and Technology, Reinventing Ihde. Springer, pp. 141–155. doi: 10.1007/978-3-030-35967-6_9.

Verbeek, P.-P. (2005) What Things Do: Philosophical reflections on technology, agency, and design. The Pennsylvania State University Press.

Verbeek, P.-P. (2006) 'Materializing Morality. Design Ethics and Technological Mediation', Science, Technology, & Human Values, 31(3), pp. 361–380.

Verbeek, P.-P. (2009) 'Let's Make Things Better: A Reply to My Readers', Human Studies, 32(2), pp. 251–261. doi: 10.1007/s10746-009-9118-0.

Verbeek, P.-P. (2010) 'Accompanying Technology: Philosophy of Technology after the Ethical Turn', Techne: Research in Philosophy and Technology, 14(1), pp. 49–54.

Verbeek, P.-P. (2011) Moralizing Technology: Understanding and Designing the Morality of Things. The University of Chicago Press.

Verbeek, P.-P. (2015) 'Beyond Interaction: A Short Introduction to Mediation Theory', Interactions, 22(3), pp. 26–31. doi: 10.1145/2751314.

Veruggio, G. (2006) EURON Roboethics Roadmap.

Veruggio, G. and Abney, K. (2014) 'Roboethics: The Applied Ethics for a New Science', in Lin, P., Abney, K., and A. Bekey, G. (eds) Robot Ethics. The Ethical and Social Implications of Robotics. Cambridge, Massachusetts: MIT Press.

Veruggio, G. and Operto, F. (2006) 'Roboethics: A bottom-up interdisciplinary discourse in the field of applied ethics in robotics', International Review of Information Ethics, 6, pp. 2–8. doi: 10.4324/9781003074991-9.

Veruggio, G., Solis, J. and Van Der Loos, M. (2011) 'Roboethics: Ethics applied to robotics', IEEE Robotics and Automation Magazine, 18(1), pp. 21–22. doi: 10.1109/MRA.2010.940149.

Wallach, W. and Allen, C. (2009) Moral Machines. Teaching Robots Right from Wrong. Oxford University Press.

Weber, M. (2012) El político y el científico. Madrid: Alianza Editorial.

Weng, Y. H. and Hirata, Y. (2018) 'Ethically Aligned Design for Assistive Robotics', in 2018 International Conference on Intelligence and Safety for Robotics. IEEE, pp. 286–290. doi: 10.1109/IISR.2018.8535889.

Willinger, I. (2019) Hi, A.I. Available at: https://www.filmin.cat/pelicula/hi-a-i.

Winner, L. (1980) 'Do Artifacts Have Politics?', Daedalus, 109(1), pp. 121–136. Available at: http://www.jstor.org/stable/20024652.

Winner, L. (2009) 'Do Artifacts Have Politics?', in Kaplan, D. M. (ed.) Readings in the Philosophy of Technology. 2nd edn. Rowman & Littlefield Publishers, Inc., pp. 251–263.

WIPO (2021) 'WIPO Technology Trends 2021: Assistive Technology'. Geneva: World Intellectual Property Organization. Available at: http://assistiveeducationaltechnology.weebly.com/assistive-technology.html#.

Yew, G. C. K. (2020) 'Trust in and Ethical Design of Carebots: The Case for Ethics of Care', International Journal of Social Robotics. Springer Netherlands, (April). doi: 10.1007/s12369-020-00653-w.

Zardiashvili, L. and Fosch-Villaronga, E. (2020) '"Oh, Dignity too?" Said the Robot: Human Dignity as the Basis for the Governance of Robotics', Minds and Machines. Springer Netherlands, 30(1), pp. 121–143. doi: 10.1007/s11023-019-09514-6.

Žižek, S. (2011) Violència. Barcelona: Editorial Empúries.

Žižek, S. (2013) The Purpose of Philosophy is to Ask the Right Questions, Big Think. Available at: https://bigthink.com/videos/the-purpose-of-philosophy-is-to-ask-the-right-questions.