Call for papers: "Intelligenza Artificiale: prospettive bioetiche, biogiuridiche e sociali"

Condotte di ricerca discutibili e irresponsabili: il ruolo di sviluppatori, annotatori e utenti di sistemi di Intelligenza Artificiale

Questionable and Irresponsible Research Practices: The Role of Developers, Annotators, and Users of Artificial Intelligence Systems

LUDOVICA MARINUCCI
ludovica.marinucci@ethics.cnr.it

AFFILIAZIONE
Centro Interdipartimentale per l'Etica e l'Integrità nella Ricerca, CNR

SOMMARIO

L'articolo si propone di analizzare alcuni casi di sistemi di Intelligenza Artificiale (IA) che hanno sollevato preoccupazioni sia nell'opinione pubblica che all'interno della comunità di ricercatori/sviluppatori di IA. Questi esempi sono in grado di mostrare le conseguenze di pratiche di ricerca 'discutibili' e 'irresponsabili' dal punto di vista non solo dei principi morali ma soprattutto degli standard professionali di ricercatori e sviluppatori. Analisi sistematiche dello stato dell'arte dei sistemi di IA, sviluppati grazie a metodologie e tecnologie molto diverse che hanno portato a output indesiderati ed esperimenti falliti, sono necessarie non solo per definire standard e codici di condotta ma anche per aumentare la consapevolezza dei ricercatori/sviluppatori di IA dei principali comportamenti irresponsabili in cui possono incorrere, anche apparentemente non gravi, comprendendone così l'impatto e le ricadute a livello individuale e sociale.

PAROLE CHIAVE

Intelligenza Artificiale

Etica della ricerca

Integrità nella ricerca

Condotte di ricerca discutibili

ABSTRACT

The article aims to analyze some cases of Artificial Intelligence (AI) systems that have raised concerns both from the public opinion and within the AI researcher/developer community. These examples are able to show the consequences of 'questionable' and 'irresponsible' research practices from the point of view not only of moral principles but above all of the professional standards of researchers and developers. Systematic analyses of the state of the art of AI systems, developed thanks to very different methodologies and technologies that have led to unwanted outputs and failed experiments, are necessary not only to define standards and codes of conduct but also to increase the awareness of AI researchers/developers of the main irresponsible behaviors they may incur, even apparently not serious, thus understanding of their impact and repercussions at individual and social level.

KEYWORDS

Artificial Intelligence

Research Ethics

Research Integrity

Questionable research practices

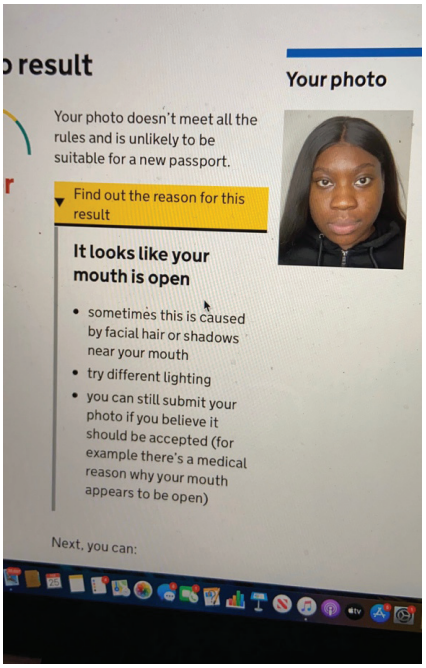
DOI: 10.53267/20240107

1. RISULTATI INDESIDERATI ED ESPERIMENTI FALLITI

L'articolo intende analizzare alcuni casi di sistemi di Intelligenza Artificiale (IA) che hanno suscitato dibattiti e preoccupazioni tanto da parte dell'opinione pubblica quanto all'interno della comunità di ricercatori/sviluppatori in ambito di IA. Tali esempi hanno lo scopo di mostrare le rilevanti conseguenze di condotte di ricerca che, seppur non definibili come *intenzionalmente* 'scorrette', come la falsificazione e il plagio, sono 'discutibili' (*questionable*) e 'irresponsabili' (*irresponsible*) dal punto di vista non solo dei principi morali (*Research Ethics*) ma soprattutto degli standard professionali (*Research Integrity*) di ricercatori e sviluppatori, secondo la nota distinzione dello storico della scienza Nicholas H. Steneck[1]. Alcuni tra gli esempi più noti sono stati riconosciuti dalla stessa comunità di ricercatori in ambito di IA, e in particolare di 'apprendimento automatico' (*machine learning*), come «an alarming red flag on our behavior as researchers and developers, since our actions can have a direct impact on society[2]». Prese di posizione di questo tipo, associate ad analisi sistematiche dello stato dell'arte di sistemi, basati su metodologie e tecnologie anche molto diverse che hanno portato ad risultati (*output*) indesiderati e a esperimenti falliti, sono il primo passo verso la consapevolezza di ricercatori/sviluppatori in ambito di IA tanto dei rischi e delle implicazioni sociali quanto delle possibili ripercussioni psicologiche che tali applicativi, a volte imposti dagli stessi governi, possono avere sugli utenti.

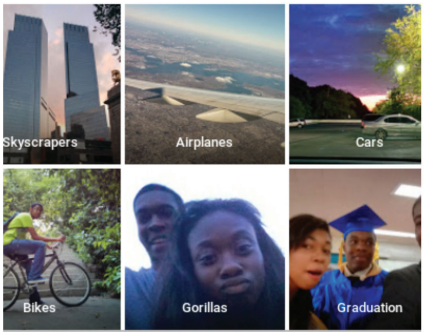
Proprio la prospettiva dell'utente, inteso come utilizzatore passivo, è rappresentata dal caso del sistema automatico per la creazione dei passaporti britannici che nel 2020 ha erroneamente riconosciuto delle labbra prominenti, caratteristiche delle persone con origini africane, come una bocca aperta.

Nel post ironico pubblicato sull'allora account Twitter (@elainebabey) dalla donna in questione, sono ancora visibili le possibili 'spiegazioni' fornite dal sistema: in particolare quelle relative alla possibile presenza di peli o di luce insufficiente sul viso inquadrato della donna mostrano come la discriminazione dell'utente sia la conseguenza di una "questionable research practice"[3], ovvero l'introduzione di 'bias' da parte dei ricercatori nelle metodologie seguite e nei risultati ottenuti e, nel caso specifico, negli strumenti informatici costruiti. Non a caso, come per gli studiosi della cognizione umana[4], il problema dei 'bias' è uno dei principali temi affrontati dai ricercatori di *machine learning*[5]. Infatti, dalle inclinazioni sociali presenti nei dati disponibili ai pregiudizi personali inseriti nel processo di creazione e rilascio di un sistema, ogni passaggio di una *pipeline* di apprendimento automatico è soggetta all'introduzione di una qualche forma di 'bias'. Già nella fase di selezione, raccolta e annotazione dei dati i ricercatori possono introdurre bias cognitivi, quali ad esempio il c.d. 'errore di campionamento' (*sampling bias*) oppure il c.d. 'favoritismo all'interno del gruppo' (*in-group favouritism*) correlato alla 'discriminazione all'esterno del gruppo' (*out-group discrimination*). Successivamente, le scelte algoritmiche da fare sul modello (*loss functions*, *regularization terms*, ecc.) possono ampliare qualsiasi 'bias' preesistente sui dati e, infine, il modo in cui sono analizzati e presentati i dati può essere deviato da bias interpretativi, come il c.d. 'pregiudizio di conferma' (*confirmation bias*) e molti altri. Questi errori a cui diamo la connotazione morale di 'pregiudizio', rappresentano la punta dell'iceberg di processi cognitivi, funzionali e definiti come "adaptive toolbox"[6], i quali restano sommersi e quasi inaccessibili, tanto per la cognizione umana che per quella artificiale. A quest'ultima, definita come una 'scatola nera' (*black-box*), è stata imposta dalle europee *Ethics Guidelines for Trustworty AI*[7] quella 'esplicazione' (*explainability*) utile non solo agli utenti ma soprattutto ai ricercatori/



sviluppatori per elaborare strategie di mitigazione dei 'bias' (*debiasing*) già nelle prime fasi di progettazione dei sistemi di IA.

La presenza di 'bias' desta particolare clamore quando essi concorrono a creare *output* discriminatori imbarazzanti, nonché controproducenti a livello economico, per le stesse aziende che producono tali sistemi a fini commerciali. Come rilevato da vari studi in ambito di 'visione artificiale' (*computer vision*), i modelli commerciali di riconoscimento facciale hanno una notevole diminuzione delle prestazioni con soggetti con la pelle più scura, soprattutto se donne[8]. Tra i primi cattivi esempi figura *Google Photo* che nel 2015 ha associato l'etichetta 'gorilla' (*gorillas*) alla foto di due persone dalla pelle scura[9]. Nonostante le scuse pubbliche, la Big Tech americana non ha provveduto davvero a risolvere il problema a livello tecnologico, optando per l'eliminazione dell'etichetta 'gorilla' e simili ('scimpanzé', 'scimmia', ecc.) anche in relazione alle immagini di quei primati.



L'esempio dimostra come algoritmi e dispositivi, anche di uso quotidiano, hanno il potenziale di diffondere e rafforzare stereotipi dannosi. Tali pregiudizi espongono alcune categorie di persone, e in particolare le donne di colore, al rischio di essere lasciate indietro nella vita economica, politica e sociale. Infatti, gli algoritmi non solo forniscono consigli sui film e prodotti da acquistare, ma sono anche sempre più utilizzati per prendere decisioni ad alto rischio, ad esempio nelle valutazioni dei pazienti[10], delle domande di prestito bancario[11], dei candidati da assumere[12] e persino delle probabilità di recidiva di un imputato[13], mostrando *output* discriminatori basati sull'etnia e sul genere. Ad oggi, nonostante i dibattitti su 'pregiudizi' ed 'equità' (*fairness*) nei sistemi di apprendimento automatico[14], i numerosi tentativi di "debiasing" sia tramite *post-processing*[15] sia direttamente durante il *training*[16], nonché gli sforzi per la creazione di grandi dataset rappresentativi di diverse etnie[17], la situazione non sembra cambiata. L'avvento dell'IA generativa, basata su grandi modelli linguistici (*Large Language Models*), solleva le stesse preoccupazioni circa il perpetuarsi di 'pregiudizi' sistemici incorporati nei dati di *pre-training*, come dimostra uno studio che esplora il potenziale pregiudizio etnico e di genere di ChatGPT: alla richiesta di valutazione di CV fittizi di candidati arabi, asiatici, afro e centroafricani, europei, americani e sudamericani le risposte discriminatorie del *chatbot* si basano ancora su un meccanismo statistico che riecheggia stereotipi sociali[18] di cui gli utenti, di qualunque tipo, devono essere resi consapevoli.

Tuttavia, prima ancora che un problema etico, si tratta qui di aspetti che hanno a che fare con l'adesione a certi standard professionali da parte di ricercatori in ambito informatico, i quali implicano strategie volte a correggere pratiche scorrette nella progettazione e valutazione dei sistemi. Visti da questa prospettiva, infatti, gli innumerevoli tentativi di mitigazione e prevenzione dei 'bias' da parte della comunità dei ricercatori/sviluppatori che utilizza tecniche di *machine learning* sembrano consistere in buone pratiche (*best practices*) riconducibili a nient'altro che 'condotte di ricerca responsabili'. Ad esempio, nella fase di *pre-training* del sistema, per garantire la diversità dei dati è consigliato selezionare e combinare *input* da più fonti di dati; mentre per ottenere un'annotazione accurata bisognerebbe non solo prevedere un team diverso rispetto a chi ha selezionato i dati ma anche ricorrere ad esperti esterni per rivedere il lavoro svolto[19]. Vale la pena sottolineare che la fase di annotazione dei dati è forse la più delicata e critica nell'orientare gli *output* di sistemi basati sull'apprendimento automatico, data la forte componente di soggettività umana nell'attività richiesta al c.d. 'annotatore', attore fondamentale del processo di cui il ricercatore/sviluppatore deve tenere conto, oltre all'utente. Nel corso degli anni, il ruolo di annotatori e revisori di annotazioni precedenti si è reso così fondamentale da ideare piattaforme, come *Amazon Mechanical Turk*, le quali erogando un compenso minimo[20], considerabile come sfruttamento in molti paesi occidentali, sollevano rilevanti criticità etiche relative al loro utilizzo all'interno di progetti di ricerca e sviluppo. A livello teorico, inoltre, uno dei problemi di ricerca più complessi e rilevanti, come testimoniano anni di tentativi per trovare metriche di 'accordo' tra

annotazioni diverse[21], sistemi di valutazione di annotazioni multiple[22] e approcci iterativi che assegnano al campione solo le annotazioni unanimi[23], è quello di raggiungere un 'consenso' tra annotatori indipendenti che elimini l'eccessiva soggettività delle annotazioni, molto evidente nel caso di concetti astratti come emozioni, sentimenti e valori[24]. A tal fine, le principali *best practices* per i ricercatori consistono sia nel definire linee guida chiare per gli annotatori sia nel creare come *gold standard* un dataset annotato in maniera ottimale per le finalità dell'applicativo.

L'assenza di un'annotazione ben bilanciata nella fase di *pre-training* del sistema è, ad avviso di chi scrive, una delle varie 'pratiche irresponsabili' che hanno portato al caso di Tay, *chatbot* basato sull'IA che fu lanciato da Microsoft sull'allora Twitter nel 2016. Dopo meno di 24 ore, Microsoft dovette terminare l'esperimento perché @TayandYou aveva iniziato a generare post inappropriati con un linguaggio giudicato razzista, sessista e antisemita.

Il ruolo attivo degli utenti è stato determinante, trattandosi di un *learning software* (LS), che dopo una prima fase di addestramento continua a imparare fino ad arrivare a cambiare il suo programma in risposta alle interazioni con utenti anche mediate e amplificate, come in questo caso, da un *social network*. Perciò, le dichiarazioni di Microsoft secondo cui «within the first 24 hours of coming online, we became aware of a coordinated effort by some users to abuse Tay's commenting skills to have Tay respond in inappropriate ways[25]», non sono state sufficienti ad attenuare le critiche sull'assenza di consapevolezza e comprensione dei rischi e possibili danni che tali tipi di tecnologie possono comportare. Gli sviluppatori di Microsoft dovevano prevedere l'alta probabilità dell'esito dell'esperimento, dato che un «LS always has this sort of vulnerability, and therefore, a developer of LS should adopt a position of expecting this behavior. The developer cannot be confident about knowing how the system will behave because of the nature of software that learns. [...] LS developers need to be more keenly aware of their ethical responsibilities[26]». Quindi, proprio per la loro peculiare natura, lo sviluppo dei LS richiede un'adesione a principi di etica e integrità nella ricerca, quali affidabilità, responsabilità e diligenza[27], anche maggiore rispetto ad altri *software* standard. Il caso di Tay mostra una vera e propria 'condotta di ricerca irresponsabile' (*irresponsible research practice*) da parte dei ricercatori/sviluppatori di Microsoft che hanno sottovalutato l'importanza della fase di progettazione e pre-addestramento del sistema per mitigare i rischi della fase sperimentale di apprendimento aperto con utenti sconosciuti e anonimi. Inoltre, il team coinvolto nell'esperimento online avrebbe dovuto monitorare più diligentemente l'evoluzione delle risposte offensive del *chatbot* in modo tale da nasconderle al pubblico. Invece, non è chiaro se la rimozione di Tay da parte di Microsoft sia stata una reazione ai tweet offensivi del chatbot o una reazione all'indignazione degli utenti verso di essi. Nel secondo caso, Microsoft non solo non avrebbe previsto questa possibilità in anticipo, ma avrebbe anche sottovalutato il problema nascente lasciando attiva e libera Tay il più a lungo possibile, ovvero fino a quando la pressione mediatica ha reso evidente la necessità di terminare l'esperimento. Qualunque sia la verità, concordiamo sul fatto che questo caso è capace di evidenziare nodi cruciali per la definizione di «appropriate professional best practice for internal processes when "releasing" LS to the general public[28]».

2. IL NECESSARIO LEGAME TRA FORMAZIONE, RICERCA E INNOVAZIONE

Questa analisi preliminare di specifiche tecnologie utilizzate per lo sviluppo di sistemi di IA, così come dei diversi attori coinvolti, è stato un primo passaggio finalizzato a evidenziare e specificare alcuni aspetti rilevanti per la definizione di norme e codici di condotta utili a ricercatori/sviluppatori in ambito informatico, che devono essere emanati da parte dei loro datori di lavoro (quali università e istituti di ricerca) e, con una visione più ampia, dai governi nazionali e sovranazionali, eventualmente anche aggiornando attuali linee guida di etica e integrità nella ricerca[29]. A tal fine, si rende necessaria da parte di gruppi multidisciplinari

di esperti una ricognizione e analisi sistematiche degli aspetti fondamentali del processo di progettazione, valutazione e distribuzione di sistemi di IA relativi tanto al tipo di tecnologie utilizzate per implementare i sistemi (*supervised machine learning*, *reinforcement learning*, ecc.) quanto al ruolo peculiare degli attori coinvolti (utenti attivi, annotatori, ecc.). Questo approccio richiede un coordinamento tra la comunità scientifica, i decisori politici e le altre parti interessate (*stakeholders*) che deve riflettersi non solo nella definizione di codici e linee guida ma anche nei bandi di finanziamento e nella valutazione dei risultati dei progetti. In tale contesto, gli stessi risultati derivanti da fondi pubblici, ad esempio dataset di qualità ad accesso aperto, devono essere diffusi e condivisi attraverso le infrastrutture di ricerca esistenti a livello nazionale ed europeo che favoriscano approcci collaborativi e di riuso di tali risultati seguendo standard di integrità nella ricerca. Tali standard mirano a salvaguardare la ricerca da distorsioni dovute a interessi economici e politici che, come abbiamo visto, sono particolarmente evidenti nei sistemi di IA sviluppati a uso commerciale da grandi aziende private. In particolare, gli esempi sopramenzionati evidenziano la necessità tanto di una maggiore attenzione e tutela verso annotatori e utenti quanto di un maggior controllo su piattaforme e strumenti collaborativi utilizzabili nelle varie fasi di addestramento e valutazione di sistemi di IA sviluppati nell'ambito di progetti di ricerca.

Un requisito fondamentale all'uso da parte dei ricercatori/sviluppatori di tali strumenti collaborativi dovrebbe essere l'erogazione di una formazione preliminare relativa non solo alle attività tecniche da svolgere ma anche alle responsabilità, all'impatto e ai rischi specifici del sistema di IA che si sta collaborando a implementare. In ambito di IA, infatti, tutti gli attori coinvolti nelle varie fasi di sviluppo dovrebbero essere formati su principi, criteri e pratiche per una condotta di ricerca responsabile, la quale «is simply conducting research in ways that fulfill the professional responsibilities of researchers, as defined by their professional organizations, the institutions for which they work and, when relevant, the government and public[30]». Tale definizione si basa sull'assunto secondo cui la ricerca scientifica vada considerata come un'attività professionale condotta da persone che hanno ricevuto una formazione specifica. Tuttavia, se possiamo dare per scontata la formazione scientifica, teorica e pratica, dei ricercatori sui contenuti peculiari del loro ambito di ricerca (nel caso specifico, quello dell'IA), in continuo aggiornamento sui più recenti sviluppi tecnologici, sembra meno evidente un'adeguata conoscenza dei principali comportamenti scorretti, anche apparentemente poco gravi, e comprensione del loro impatto e ripercussioni sociali. Una 'formazione culturale e civile' su principi morali e codici professionali che guidino ricercatori, annotatori, utenti, ecc., non solo su *cosa* dovrebbero fare o meno, ma anche su *come* dovrebbero farlo, eviterebbe buona parte degli esempi di 'risultati indesiderati' qui considerati. Il dibattito e l'attenzione mediatica suscitati, e che continuano ciclicamente a suscitare, reclamano la pressante necessità di una sensibilizzazione volta a far percepire tali questioni non come meri adempimenti burocratici per pubblicazioni o fondi di ricerca, ma come requisiti essenziali per ottenere risultati che puntino a costruire una "società buona"[31] o, almeno, a evitare danni tanto individuali quanto sociali e pubblici, soprattutto in termini di investimenti sprecati o di un indebolimento della fiducia nei riguardi dei sistemi di IA sviluppati e, per estensione, dei loro ricercatori/sviluppatori.

NOTE

1. N. H. Steneck, "Fostering integrity in research: Definitions, current knowledge, and future directions", *Science and engineering ethics* 12 (2006), 53-74, https://doi.org/10.1007/PL00022268

2. C. Laranjeira, V. Fernandes Mota, and J. A. dos Santos, "Machine Learning Bias in Computer Vision: Why do I have to care?", in *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images*, IEEE (2021), p. 1.

3. N. H. Steneck, 2021, *cit*.

4. A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," in *Rationality in action: Contemporary approaches*, ed. P. K. Moser (Cambridge University Press, 1990), 171–188 [Reprinted from *Science* 185 (1974), 1124-31]

5. N. Mehrabi et al., "A survey on bias and fairness in machine learning," *ACM computing surveys - CSUR*

54, no. 6 (2021), 1-35, https://doi.org/10.1145/3457607

6. G. Gigerenzer and H. Brighton, "Homo heuristicus: Why biased minds make better inferences," *Topics in cognitive science* 1, no. 1 (2009), 107-143, https://doi.org/10.1111/j.1756-8765.2008.01006.x

7. HLEG - High-Level Expert Group on AI, "Ethics Guidelines for Trustworthy AI," (2019), https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

8. J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in P*roceedings of the 1st Conference on Fairness, Accountability and Transparency*, vol. 81, (New York, NY, USA: PMLR, 23–24 Feb 2018), 77–91, http://proceedings.mlr.press/v81/buolamwini18a.html

9. BBC News, "Google apologises for Photos app's racist blunder," July 1, 2015, https://www.bbc.com/news/technology-33347866

10. Z. Obermeyer, B. Powers, C. Vogeli et al., "Dissecting racial bias in an algorithm used to manage the health of populations," *Science* 366, no. 6464 (2019), 447–453, https://www.science.org/doi/10.1126/science.aax2342

11. A. Mukerjee, R. Biswas, K Deb et al., "Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management," *International Transactions in operational research* 9, no. 5 (2002), 583-597, https://doi.org/10.1111/1475-3995.00375

12. A. Peng, B. Nushi, E. Kıcıman, et al., "What you see is what you get? the impact of representation criteria on human bias in hiring," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7 (2019), 125-134, https://doi.org/10.1609/hcomp.v7i1.5281

13. J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science advances* 4, no. 1 (2018), eaao5580, https://www.science.org/doi/full/10.1126/sciadv.aao5580

14. N. Mehrabi et al., 2021, *cit.*

15. T. Bolukbasi, K.W. Chang, J.Y. Zou et al., "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in NeurIPS Proceedings of *Advances in neural information processing systems* 29 (2016), 4349–4357, https://bit.ly/4h2gX87

16. J. Zhao, T. Wang, M. Yatskar et al., "Gender bias in contextualized word embeddings," (2019), https://doi.org/10.48550/arXiv.1904.0331

17. H. J. Ryu, M. Mitchell, H. Adam, "Inclusivefacenet: Improving face attribute detection with race and gender diversity," in *Proceedings of Workshop on Fairness, Accountability, and Transparency in Machine Learning,* (FAT/ML 2018), https://www.fatml.org/media/documents/inclusive_facenet_zOOhwRN.pdf

18. L. Lippens, "Computer says 'no': Exploring systemic bias in ChatGPT using an audit approach," *Computers in Human Behavior: Artificial Humans* 2, no. 1 (2024), 100054, https://doi.org/10.1016/j.chbah.2024.100054

19. B. Cowgill, F. Dell'Acqua, S. Deng et al., "Biased programmers? or biased data? a field experiment in operationalizing ai ethics," in ACM Conference on Economics and Computation, 2020, pp. 679-681. http://dx.doi.org/10.2139/ssrn.3615404

20. K. Fort, G. Adda and K.B. Cohen, "Amazon Mechanical Turk: Gold mine or coal mine?," *Computational Linguistics* 37, no. 2 (2011), 413–420, https://doi.org/10.1162/COLI_a_00057

21. M. Fuoli and C. Hommerberg, "Optimising transparency, reliability and replicability: Annotation principles and inter-coder agreement in the quantification of evaluative expressions," *Corpora* 10, no. 3 (2015), 315-349, https://doi.org/10.3366/cor.2015.0080

22. F. Rodrigues, F. Pereira and B. Ribeiro, "Learning from multiple annotators: distinguishing good from random labelers," *Pattern Recognition Letters* 34, no. 12 (2013), 1428–1436, https://doi.org/10.1016/j.patrec.2013.05.012

23. D. M. Iraola and A. J. Yepes, "Single versus multiple annotation for named entity recognition of mutations" (2021), https://doi.org/10.48550/arXiv.2101.07450

24. J. Hoover, G. Portillo-Wightman, L. Yeh et al., "Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment," *Social Psychological*

and Personality Science 11, no. 8 (2020), 1057-1071, https://doi.org/10.1177/1948550619876629

25. P. Lee (2016), "Learning from Tay's introduction," Official Microsoft Blog, https://bit.ly/49Jqj5X

26. M.J. Wolf, K.W. Miller, F.S. Grodzinsky, "Why we should have seen that coming: comments on Microsoft's Tay "experiment, and wider implications," *ACM SIGCAS Computers and Society* 47, no. 3 (2017), p. 3, https://doi.org/10.29297/orbit.v1i2.49

27. Commissione per l'Etica e l'Integrità nella Ricerca del CNR, "Linee guida per l'integrità nella ricerca," 2019, https://bit.ly/3P2B7CU

28. M.J. Wolf et al., 2017, *cit.*, p. 6.

29. ALLEA, "The European Code of Conduct for Research Integrity," Revised Edition 2023, Berlin, DOI 10.26356/ECOC, https://bit.ly/4ilkCcF

30. N. H. Steneck, 2021, *cit.*, p. 55.

31. L. Floridi, J. Cowls, M. Beltrametti, et al., "AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations," *Minds and machines* 28 (2018), 689–707, https://doi.org/10.1007/s11023-018-9482-5