AI and Ethics
https://doi.org/10.1007/s43681-024-00632-1

EDITORIAL

Exploring the mutations of society in the era of generative AI

Hubert Etienne[1] · Brent Mittelstadt[2] · Rob Reich[3] · John Basl[4] · Jeff Behrends[5] · Dominique Lestel[6] · Chloé Bakalar[7] · Geoff Keeling[8] · Giada Pistilli[9] · Marta Cantero Gamito[10]

Received: 26 November 2024 / Accepted: 26 November 2024
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2024

In the past two years, Generative AI has acquired a central position in the public space to become a key concern for governments when revising their national strategies as much as for private companies trying to integrate it into their business model. New threats are being considered, such as impersonation techniques, data poisoning and large disinformation campaigns to destabilize political regimes, and new capabilities are emerging with the democratization of AI systems' use through natural language. As AI systems' capabilities grow, new questions also emerge regarding the creative process and property rights.

Beyond the business opportunities it brings about, the large adoption of generative AI leads us to consider some fundamental questions for our societies: How should we define data ownership and compensate those who produce the data upon which LLMs are trained? To what extent should foundational models be open to allow citizen control through compliance audits without benefiting malicious actors? How may generative AI impact the job market overall and among social classes? Will it result in universal-basic-income-based societies powered by artificial workers? What types of interactions may humans develop with artificial agents and how is it likely to impact their cognitive capacities? What are the skills they will lose and those they will develop? Will artificial agents be granted moral agency and may progress towards AGI trigger geopolitical tensions?

On June 6th and 7th 2024, the Paris Conference on AI & Digital Ethics second edition gathered leading academic and industry researchers at Sciences Po, Paris, to address such questions collaboratively and transdisciplinary. This topical collection contains eight papers presented at the conference and resulting from a rigorous selection among sixty-four proposed contributions. Together, they underscore the profound changes our societies may undergo with the massive adoption of generative AI solutions to open new research domains and drive the responsible development of such solutions by informing their actors.

Hubert Etienne– Lead editor

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

✉ Hubert Etienne
hubert.etienne@quintessenceai.com

Brent Mittelstadt
brent.mittelstadt@oii.ox.ac.uk

Rob Reich
reich@stanford.edu

John Basl
j.basl@northeastern.edu

Jeff Behrends
jbehrends@fas.harvard.edu

Dominique Lestel
lestel@ens.fr

Chloé Bakalar
cbakalar@meta.com

Geoff Keeling
gkeeling@google.com

Giada Pistilli
giada@huggingface.com

Marta Cantero Gamito
marta.cantero@eui.eu

[1] Quintessence AI, New York, USA
[2] University of Oxford, Oxford, UK
[3] Stanford University, Stanford, USA
[4] Northeastern University, Boston, USA
[5] Harvard University, Cambridge, USA
[6] Ecole Normale Supérieure, Paris, France
[7] Meta, New York, USA
[8] Google, London, UK
[9] Hugging Face, Paris, France
[10] European University Institute, Fiesole, Italy

Published online: 13 January 2025

AI and Ethics
https://doi.org/10.1007/s43681-024-00629-w

ORIGINAL RESEARCH

AI to renew public employment services? Explanation and trust of domain experts

Thomas Souverain[1]

Received: 1 July 2024 / Accepted: 16 November 2024
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2024

Abstract
It is often assumed in explainable AI (XAI) literature that explaining AI predictions will enhance trust of users. To bridge this research gap, we explored trust in XAI on public policies. The French Employment Agency deploys neural networks since 2021 to help job counsellors reject the illegal employment offers. Digging into that case, we adopted philosophical lens on trust in AI which is also compatible with measurements, on demonstrated and perceived trust. We performed a three-months experimental study, joining sociological and psychological methods: Qualitative (S1): Relying on sociological field work methods, we conducted 1 h semi-structured interviews with job counsellors. On 5 regional agencies, we asked 18 counsellors to describe their work practices with AI warnings. Quantitative (S2): Having gathered agents' perceptions, we quantified the reasons to trust AI. We administered a questionnaire, comparing three homogeneous cohorts of 100 counsellors each with different information on AI. We tested the impact of two local XAI, general rule and counterfactual rewording. Our survey provided empirical evidence for the link between XAI and trust, but it also stressed that XAI supports differently appeal to rationality. The rule helps advisors to be sure that criteria motivating AI predictions comply with the law, whereas counterfactual raises doubts on the offer's quality. Whereas XAI enhanced both demonstrated and perceived trust, our study also revealed limits to full adoption, based on profiles of experts. XAI could more efficiently trigger trust, but only when addressing personal beliefs, or rearranging work conditions to let experts the time to understand AI.

Keywords Artificial intelligence (AI) · Trust · Explainable AI (XAI) · Experimentation · Public policies · Profiles

1 Introduction

1.1 Public policies in private flows: AI investments, latent beliefs

Affected by growing burdens of debt, as critics against slow and opaque bureaucracy, the public sphere draws inspiration from private sectors where Artificial Intelligence (AI) seems to address problems of quickness, efficiency and rentability of services [16].

While public polices deploy machine-learning models from justice to fiscal administration, they inherit from the design and governance methods of private forefront companies, but also from some discourses on AI. Facing the opacity of machine-learning models due to big data, complex architectures and heavy training processes [11, 15, 25], the large panel of explainability techniques is also subject to implicit assumptions. One latent belief is the following: "Explainability fosters trust in AI" [4] of public agents.

Yet, there is a lack of field works actually testing if explainable AI (XAI), or the field mostly represented by computer scientists to design techniques to simplify or trace back AI structured and outcomes [15], will be sufficient to promote trust in AI [16]. The public services are also influenced by this dark screen in XAI research, all so the more they bear the additional duty to consider their users impacted by AI as citizens.

✉ Thomas Souverain
thomas.souverain@ens.psl.eu

[1] Institut Jean Nicod, École Normale Supérieure de Paris (CNRS, ENS-PSL, EHESS) & France Travail AI Department, Paris, France

Published online: 09 January 2025

1.2 Our use case—AI classifier warning for illegal job offers

In this article, we propose an interdisciplinary method to test the link between trust in AI outcomes and XAI, as to anchor this method on rigorous notions of trust.

We led qualitative and quantitative surveys on public agents which were confronted with AI. Based on external recruiters demands, the French Employment Agency (France Travail) publishes 200 000 new job offers by month. Nothing illegal must be published, from "cleaning lady" (gender discrimination) to "bring your own vehicle" (for non-licensed professions). Yet, it was most challenging for the job counsellors to review the 200 000 offers for the 27 illegality cases. Hence, a Natural Language Processing (NLP) classifier based on neural networks was internally built and deployed in 2021, triggering alerts when a sentence was predicted to contain illegal content. On this case which had been deployed for two years, we had rich material to investigate if XAI actually supports trust of domain-experts.

2 Our method: from definitions to measures of "trust" in XAI

To measure if XAI enhances trust in AI, we must begin with a delineation of trust which is both adequate for our public services case, and precise enough to enable measures. Starting with a baseline definition is necessary to test *indicators* of trust in our experiment.

2.1 Trust as reliance

Discussions on XAI generally agree on reliance to be at least a necessary condition for trust in AI. The "trust-fall game" illustrates this dimension of trust in machines [20]. For those who never plaid it in the schoolyard, this game consists in someone (the trustor) which allows oneself to fall backwards, believing that other people—supposed friends until now (the trustee)—will catch him or her.

In the trust-fall game, the pupil precisely *relies on* her classmates, in the sense that she accepts to be dependent on the intervention of others in a context of uncertainty. Reliance is constituted by an *action* of delegation (letting herself fall, Cf. Fig. 1 in green) based on a *belief* on uncertain others' operations (the classmates will catch her, Cf. Fig. 1 in red).

The duality between action and belief raises problems to measure reliance. As [21] observes, empirical studies on trust in AI generally collect *perceptions of the trustees on their trust in AI*. Most quantitative studies ask respondents on their degree of agreement with assertions such as "I am confident in the tool" [9] or "I trust the recommendations by algorithms-driven services" [27]. Yet, in situ experiments show inconsistency between peoples' rating of their own trust and how they actually behave [20]. Qualitative studies like interviews escape the abstraction risks of positioning

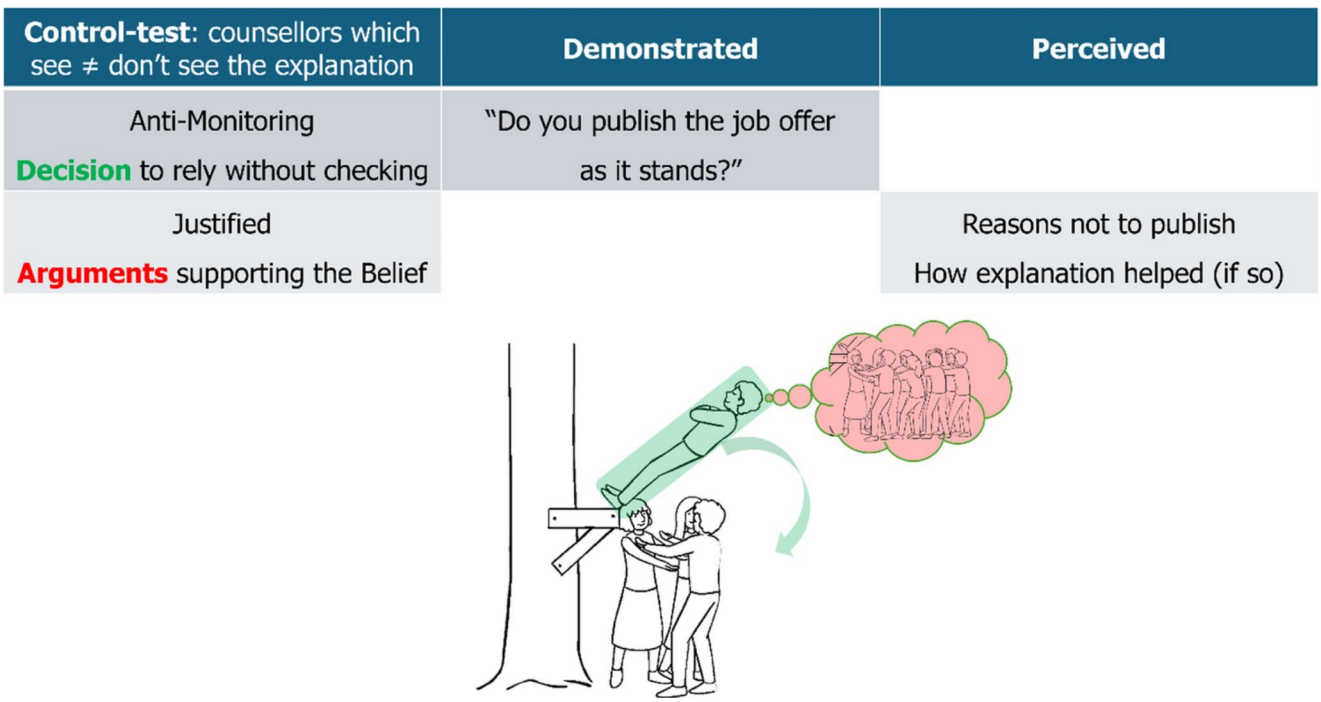


Fig. 1 Schematic Introduction to Trust as Reliance on AI—the trust-fall game

on a trust-scale, so far as they investigate the operational conditions of the trustors [11]. However, they also introduce declarative and presentation biases in front of the researcher [21].

Thus, besides "perceived" trust, [21] rightly observes that the "actual" or "demonstrated" one falls beyond the scope of these studies. Say, the *decision* to let oneself fall at the very moment the trustor faces up the AI outcome, cannot be observed when reconstructing one's level of trust or the reasons to rely on AI. To take into account demonstrated and perceived trust, we must dig further into the conditions for relying on AI outcomes.

Helpfully, among the rare attempts to define trust in AI on concrete situations, some discussions on medical devices introduced the frame of "computational reliabilism" [4]. Two conceptual features identified by [4] to account for trust in electronic health records will sharpen our criteria for reliability.

2.2 Demonstrated trust—no monitoring when XAI is present (control-test)

The first practical condition is "anti-monitoring". The agent is confident enough to accept the result of an AI system, without additional checking. It helps us complement our empirical criterion for demonstrated trust. In our case of interest, the recruiting expert facing an AI alert of illegality can choose to delete or modify it according to the AI outcome, or to take the opposite decision—publishing the offer despite the alert.

However, in order to make anti-monitoring *demonstrating* trust, we need to introduce a sense of causality in our experiment. It must indeed be proved that trust in AI is not incidental, but instead, triggered by XAI. Therefore, we introduce a control-test setting, conforming to an *interventional* approach of causality [2, 24]. In order to demonstrate the effect of trust in XAI on the same job offer, some job counsellors must see the additional explanation (test) whereas other with same demographics only see the job offer (control). Trust would hence be *demonstrated* when everything else being equal, the test-advisors which align with AI outcome are in growing number compared to control-advisor deciding not to publish.

Based on interventional causality, our definition of trust as reliance comes with an empirical translation, that is, domain experts acting as AI outcomes recommend *when an explanation is present* (compared to when it is not).

2.3 Perceived trust—epistemic warranties

As baseline definitions become clearer in our job offering context, on which beliefs (Cf. Fig. 1, red part) is this demonstrated trust (Cf. Fig. 1, green part) based? In the trust-fall game, the agent has in mind that a large part of schoolmates is going to catch her. However, in real cases as the professional use of AI to detect illegal job advertising, the internal model of agents is not directly accessible.

One way of investigating if their trust in AI is "justified" by the supplementary material of XAI [4] stands on the side of *perceived* trust. Indeed, if the reasons expressed by a job counsellor to trust an AI alert fit with the information brought by XAI, we can assert that trust is justified.

Thereby, we provide a complementary frame to think about trust in XAI, in our job advertising context. As [21], we distinguish between the trust which is perceived, and the one which is demonstrated by reliance due to XAI. We supplement this vision with the one of [4], to specify demonstrated trust as "anti-monitoring" and perceived trust as indicating if trust is "justified". Entangling the two conceptual frames on XAI and trust in AI predictions, we added causal interventions to make them operational for empirical measurement (Cf. Fig. 1). We divided the job counsellors in control-test samples, enabling to know if the intervention of XAI causally triggers more (or less) decisions to rely on AI, and at the same time, to investigate if XAI intervention changes the beliefs of the agents.

2.4 Which explanations to test?

To precise our empirical settings, we must introduce how we treated "explanation" of AI. Previous work on the same use case provided primer insights, based on human–computer interface [10]. As the users of AI investigating illegality are "non-experts" on Computer Science [3], any technical details on how the neural networks trained or proceed to reach a prediction would be useless, or indirectly useful. In spite of this, explanatory representation of AI has been proven to be useful when it conforms to the *domain-expertise* of job counsellors. Based on 30 mn survey on a panel of 14 recruiting experts, in which participants are asked to click on their preferred explanation in a pair, [10] identifies explanations in natural language which are prevalent. We retain them to be our XAI supports, and provide a hierarchy:

- Minimal Explanation (Control)—general theme: inside the job offer, the sentence on which the neural networks based its classification of illegality is simply extracted (highlighted on Fig. 2). Take the schematic example of Fig. 2, "Seeks for feminine baby-sitter…" The classification output (theme of the alert, here "Gender Discrimination") is printed while the sentence raising this NLP classification is highlighted. Corresponding to the interface jobs counsellors are used to since 2021, we take this minimal explanation as baseline.

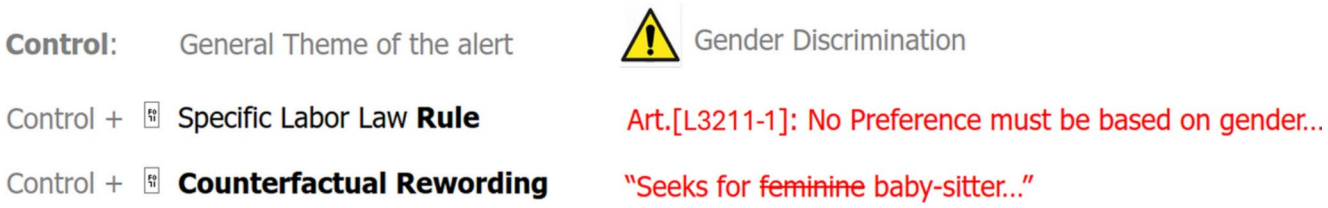


Fig. 2 Control, Rule and Counterfactual Example of the three types of information shown to counsellors

Control can be seen as a minimal explanation, in the restricted sense that it informs on the output and extracts the relevant sentence. Critics might also consider this is only a description of the "what" of AI output, whereas explanation requires to answer "why-questions" [15]. To challenge this description (or minimal explanation), we select two "why-explanations" mostly used in XAI literature for domain experts to understand local predictions [15], approximated rule [6, 25] and counterfactual example [12, 22]:

- Control + Specific Labour Law Rule: the job counsellor accesses the specific *rule* which motivates the NLP classification with the highest probability. This rule is extracted from labour law, for instance, Article L1132-1 which explicitly forbids gender-based selection of candidates.[1]
- Control + Counterfactual Rewording. The rule instructs on the general rationality subsuming the particular prediction, answering questions like "How to still be this" in [15] survey. Complementarily, counterfactual examples appeal to domain-expertise to answer "Why not" queries [15]. We generated them with slight textual changes which, everything else being equal, make the AI prediction change from illegality to legality of the job offer. For instance in Fig. 2, the counterfactual example simply wipes out "feminine" in the previous text "Seeks for feminine baby-sitter".

The cases among the 200 000 monthly job offers are of course more complicated than Fig. 2, which makes reconstructions of the specific rationale and minimal rewording relevant to identify the validity and the limits of the AI model.

We note that these explanations connect with rich philosophical piece of work on the epistemology of trust [18]. Among philosophers claiming that trust can be grounded on reasons other than the trustee's credence, internalists generally point that trust is warranted when the trustor has enough information to *rightly complete* one's mental model on what the trustee actually does [5]. In our case, rule and rewording can be considered as "processes that tend to produce accurate representation of the world" [5], as they involve the expertise of the job counsellors.

2.5 Joining qualitative and quantitative methods

Now we clarified our baseline acceptations of "trust" and "explanation" to be tested on AI illegality classification, we can give more insight on our three-month study.

- Qualitative(S1): Relying on sociological field work methods [17, 28], we conducted 1 h semi-structured interviews with job counsellors in their agencies. Each interview was precedented by a day of observation of the counsellors' practices. On 5 regional agencies, we asked 18 counsellors to describe their work practices with AI warnings. The 5 regional agencies were representative of the French employment areas, from metropolitan city to rural zones.[2]
- Quantitative(S2): Based on the agents' perceptions, we quantified the reasons to trust or distrust AI [11]. With specialists of legality in the Agency, we extracted a set of 21 job offers whose themes were representative of the aggregate volume sawn by counsellors, and labelled them according to their certified legality. We reconstituted the *most realistic* presentation of job offers as possible: their number was representative of daily volumes handled by counsellors, they were job offers extracted from real AI alerts (issued in October 2023), and the counsellors were already used to our interface.[3] We administered a questionnaire comparing 3 homogeneous cohorts of 100 counsellors, each advisor seeing the same offers in the same order, though with different information to allow for causal interpretation (XAI or control) [24].[4] Based on reasons expressed in (S1),

[1] More generally, this Article thwarts any type of job discrimination based on gender, but also on physical appearance, ethnic background, disability or beliefs. Cf. French Labour Law—Code du Travail, Article L1132-1.

[2] Identifying data of the counsellors were carefully masked (e.g. names were changed). As for the options expressed in questionnaire (S2), we translated the verbatim reports of (S1) from French.

[3] On the prior presentation of the experiment (S2) to the job counsellors, which connects with their expertise and usual AI alerts' interfaces, see Fig. 8.

[4] See Fig. 10 on the random and even distribution of explanations, and the contents of all 21 job offers.

the questionnaire gathered their decisions to publish and why, then if explanation was useful and why.[5]

At each stage of (S1) and (S2), we controlled for demographics of the advisors (year of arrival at the Agency, easiness with numerical tools, volume of job offers). On both studies, none of this factor came into significant play except the year of arrival (Cf. Sect. 4.2).

3 Results of the survey (S2)

3.1 Demonstrated trust: XAI supports increase anti-monitoring, with different roles

Let us begin with demonstrated trust. Among domain-experts without AI background in the survey (S2), adding XAI information on the local prediction generally increased trust reflected by reliance. On the 18 job offers where batches of 100 counsellors randomly saw rule, rewording or nothing more than the alert sentence, adding a "why-explanation" makes the counsellors following AI to reject the offer increase of +3% for the rule and +5% for the counterfactual rewording.

However, this incremental increase at the overall level masks the impactful roles of rule and rewording for different categories of job offers, which turned out to be very different. With the help of the Law department of the French Employment Agency, we labelled the job offers according to their *juridical* legality: "surely legal"; "debatable"; and "surely illegal". This experimental grid was key to understanding the different parts xAI supports plaid in trusting the alerts of illegality.

3.1.1 Rule in illegal cases: slight but significant increase of domain-experts relying on AI

Rule stands out where the legal experts unanimously stamped job advertisements to be illegal.

On these "surely illegal" job offers, the rule proves to increase the decision not to publish of 5% (Cf. Fig. 3). This might appear as a marginal gain. However, given the 35,000 job offers which monthly raise AI warnings, this would actually represent a decrease of 1500 unreliable job offers by month. More, this rise in anti-monitoring proves to be significant (p-value = 0.03, binomial two-sided method). The rule indeed increases the decisions not to publish in 14 of the 18 cases where the three explanations appear.

On the contrary, the counterfactual rewording (C) does not raise such attention. Over all the surely illegal job offers, the mean percentage of counsellors which follow AI to reject the content is the same for counsellors which are supplied with a counterfactual rewording (C) than for those in (A) who don't see additional explanations (70%, Cf. Fig. 3).

3.1.2 Counterfactual for unclear cases

On this observation of illegal cases where rewording does not trigger more advisors to rely on AI, we could have presumed that this type of why-explanation had no practical effect.

Yet, focusing on the employment opportunities where AI illegality classification was at least questionable (i.e. labelled by our law experts as "surely legal" or "debatable"), we were surprised to discover the *symmetric effect of why-explanations* on demonstrated trust.

Except on job offer n°15 where they reach the same level, the counterfactual local explanation (C) systematically increases the share of public advisors rejecting the offer, with a mean gain of +12% compared to (A) control (Cf. Fig. 4). This time, it is the rule (B) which does not bear significant effect on reliance, half of the time increasing and the other half decreasing anti-monitoring of job counsellors.

3.2 Perceived trust—rule to increase certainty, counterfactual raising reasonable doubts

The composition of impacted job postings confirms that XAI supports lead to different reasons for rejecting offers:

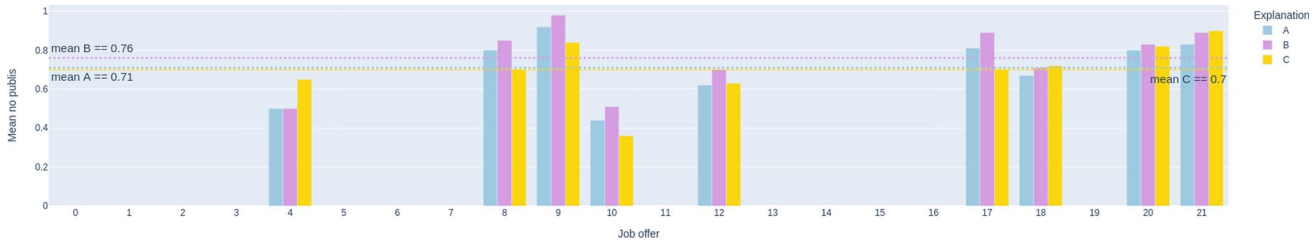


Fig. 3 Mean of counsellors which decide not to publish for every illegal job offer, depending on seen explanation A (control)/B (rule)/C (counterfactual)

[5] See Fig. 9 for the complete steps of (S2), the range of answers and the pathways between them.

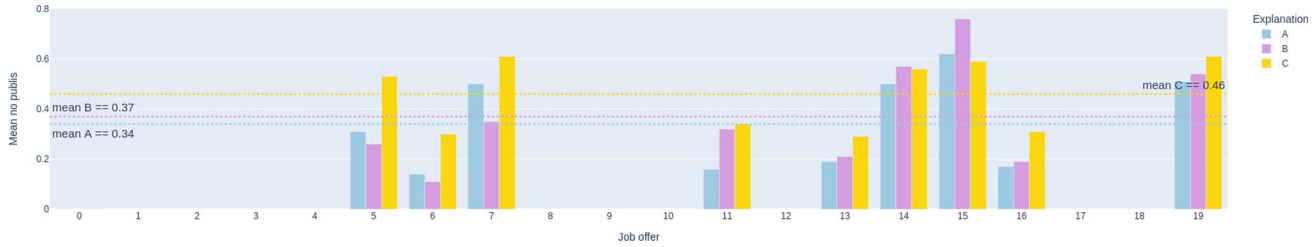


Fig. 4 Mean of counsellors which decide not to publish for every potentially or certainly legal job offer, depending on seen explanation A (control)/B (rule)/C (counterfactual)

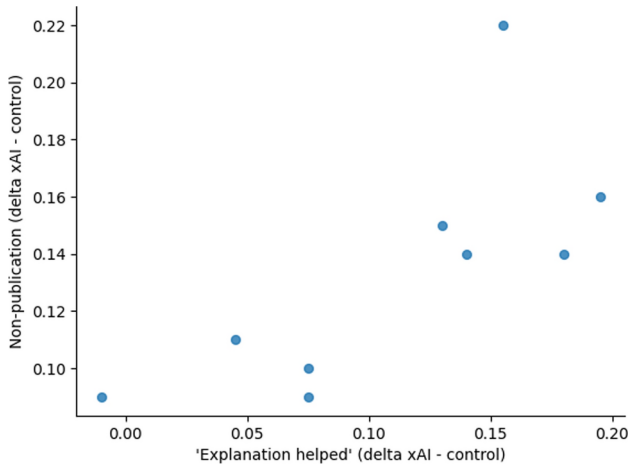


Fig. 5 Correlations between the decision and the declaration that "explanation helped", On the top 10 job offers where XAI increases non-publication

- The rule affects *certainty* on job offers. Given cases where most advisors already considered the offer as illegal, it anchors this certainty. The offers where the rule increases non-publication over 5% deal with the formulation of short-term contracts, gender identification, or clear cases where a driving-license shall not be required (e.g. in downtown).
- The counterfactual *puts in question* the reverse belief. That's why its main effect, is to raise the doubt over 15% compared to other counsellors deciding not to publish. Indeed, counterfactual mostly impacts the unclear cases, e.g. jobs on the brink of licensed professions, where the context might or not allow to use a driving-license.

3.2.1 In case it impacts decisions, XAI is perceived as trustful for appropriate reasons

The trust demonstrated by the counsellors, as their reasons for it, echo the information on AI predictions added by XAI. Though, do the domain-experts regard their heightened confidence as it is *due to the explanation*?

(S2) takes one further step toward perceived trust, asking if explanation helped and if so, why.

When rule or rewording impacts non-publication, the counsellors accordingly acknowledge that the explanation helped them more than confronted with control explanation. On the top 10 job opportunities where XAI heightens non-publication compared to control, a significant positive relationship connects the increase in counsellors which decide not to publish, and the increase in counsellors declaring that "explanation helped" (Pearson's R correlation of 0.73, Cf. Fig. 5).

Delving into the reasons for it, on the offers were the rule increased non-publication of more than + 5%, "explaining the reason behind illegality to the hiring manager" grew from 29% (control) to 42% among the counsellors exposed to the rule. For the job offers where the counterfactual has similar incidence, 14% more experts seeing the counterfactual asserted that explanation helped them "propose the hiring manager to change the offer", when compared to the experts faced with rule or control.

This supports our interpretation that beyond reinforcing demonstrated trust in complementary situations, XAI material is perceived as helpful, and for adequate reasons. The labour law rule stresses on the *limits* of legality, thus, providing counsellors with further supports to explain to hiring companies why their current offer is illegal. On the other hand, the counterfactual instance opens the mind of the domain experts to a *minimal change* which enhances the employment posting, hence, equipping them to propose a change to the hiring manager.

3.2.2 Obstacles against the effects of XAI on trust?

From these observations of complementary impact of XAI why-supports, however, claiming that bringing XAI to domain experts will systematically increase trust would be overconfident.

Indeed, among the counsellors which did not publish the offers despite they were surely illegal, the belief of their legality is enrooted despite the alert and the why-explanation.

Based on the reasons to publish in the reverse direction to the AI recommendations, four were identified during (S1) interviews, then proposed to counsellors which published despite the alert in (S2): the counsellors could (not

exclusively) argue the job opportunity to be "legal" anyway. They could also defend that the information subject to AI warning is in fact necessary to make the "candidates" fully aware of their eventual job conditions; or to enable "recruiters" to select only skilled applicants. Finally, they could simply state that due to present work conditions, they don't have "time" enough to glance at the offer.

The box plot of Fig. 6 compares on surely illegal job offers, the reasons for not publishing between counsellors which saw the rule (violet bars) and counsellors which didn't see XAI information (control, blue bars). Coherently with what we observed above, introducing a labour law rule slightly decreases the counsellors which declare to be sure the offer is legal. However, among the counsellors who don't publish, 6/10 are still convinced of the offer's legality (Cf. Fig. 6, left part). More, even for counsellors seeing the justified illegality rule, the "legality" of current job posting remains the major reason for not relying on AI.

Fig. 6 shows a ceiling effect, meaning that some counsellors persist in their own belief against AI outcomes, even when they are provided with evidence for the contrary.

4 Qualitative support of (S1)—XAI fostering trust depends on profiles

(S1) helps better understand this effect. The 18 interviews in (S1) led us to establish four profiles or advisors, based on ideal-type method crossed with demographics [17].

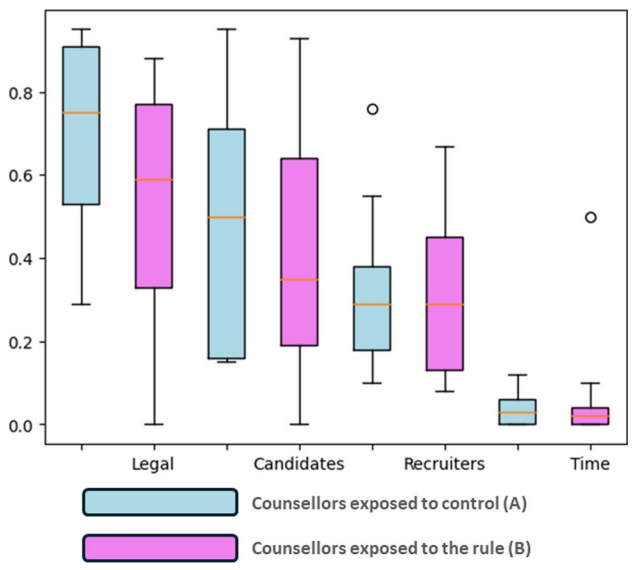


Fig. 6 Reasons expressed by the counsellors to publish despite the AI outcome, when job illegality is certified: without XAI information/when rule is present

4.1 "Heroes" and "experienced" effects: resistance where XAI does not lead to reliance

Two profiles give qualitative understanding of the resistance marked by some domain experts to trust AI, due to their *perceptions of legality* which remain unchanged. Let us begin with a verbatim report:

> "*Good shape*", yeah, they don't have the right to write it, but anyway *it is better to let the applicants know* (Sandrine. Emphasis is ours.)

Sandrine joined the French Employment Agency more than ten years ago. She works on a sparse rural area with only one urban center, where the job carriers mostly deal with hospital and building trade. Interestingly, at this moment of the interview she refers to a job offer she is handling right now. Besides, she *knows* that the posting violates labour law.[6] Sandrine nonetheless takes the decision to let the wording "Good shape" (legally discriminating). Her argument is that it is here more important to "let the applicants know" than to align with labour obligations. As she illustrates further, a candidate suffering back pain could, without this indication, come and be hired, "dramatically deteriorating one's state of health".

As Sandrine, "Heroes" profiles put at the forefront their moral beliefs, developing strong empathy for the applicants. Congruent interviews devoid significant attention to avoid applicants "encountering difficulties" and "disillusions also for the candidate which says: 'when was I told it was inaccessible?'" (Patricia, Esther). Heroes of (S1) provided qualitative understanding for why "candidates need to be aware of the information" remains high as an argument for non-publication in (S2), even when the illegality is recognized due to AI output and labour rule. Almost 4/10 counsellors express it to be a reason for publishing despite AI (Cf. Fig. 6 "Candidates", violet bar).

Beside Heroes, another profile constituted in (S2) qualitatively explains resistance to trust AI in front of justified XAI. "Experienced" share demographics with Heroes. They mostly arrived at the French Employment Agency more than ten years ago, with prior experience in recruiting or human resources.

The Experienced show blatant disregard for the "robot" which would merely rehearse what they know, and more, which could be counterproductive because it ignores the

[6] Even in construction industry, it is considered as illegal to draw differences between *job applicants* on health status. Cf. French Labour Law—Code du Travail, Article L1132-1. Only an accredited occupational doctor, not the firm, has jurisdiction to recommend specific treatments according to the mandatory medical checkup– which takes place *after* hiring. Cf. Code du Travail, Article L1133-3.

context (Marie, Jade). When they take the opposite side than AI, these profiles invoke their *experience* which would better take into account the context than the "blind" machine. On a concurrent mention of short-term and permanent contracts, Georges argue that interim agencies can need and therefore mention both, to decide after the employment interview: "I ignore the alert, *if I consider it is not justified*." (we emphasise).

Experienced bind resistance against AI because they discredit the "machine" which would automatically exclude skills required in the supplied content ("Recruiters" in Fig. 6). As the demographics of (S1) point, we also note in (S2) that Heroes and Experienced effects combine as motives against reliance on AI (strong positive Pearson-R correlation of 83% on all the 21 job offers).

4.2 XAI fosters reliance of good students, calling for their rationality

Two further effects were identified with (S1) interviews. Mostly arrived at the Employment Agency less than three years ago, the Good Students carefully align with the legal handbook provided to job counsellors. They absolutely want to "avoid discriminating words" (Ambre), systematically associating AI warnings with discrimination. Good Students are "afraid of doing wrong" (Dona), hence they don't allow any exception:

> 'Public interest', it's in my veins! Cause if we let it an illegal job offer pass once, we let it pass for all. (Jade)

In (S1), Good Students often use the lexical field of duty. But on the contrary to Heroes, they invoke their mission of applying the law in an equal manner, rather questioning their own beliefs than AI alerts. When the Good Student effect is triggered, the counsellors dedicate efforts to understand and explain AI outcomes, stressing on pedagogy to explain to recruiters why their offers have been rejected (Linda, Clara).

4.3 The "time-pressured"

> "There were 40 job offers, plus 80 offers of the day before… So we do our best. *But in the meantime, the phone rings, there are new offers* on the platform… *we miss out*…" (Ambre. Emphasis is ours.)

Ambre is not the only one who claims that she is restrained to do her job by the working structure. Time-Pressured repeat they often have "the head to the grindstone" (Linda, Sandrine, Francis), which make them "flitting" (Sandrine, Ambre, Linda). They share large consensus on the reasons for this loss in attention. Since 2021, NLP introducing new filtering was supposed to make domain-experts save time. Depending on regional agencies, their schedule was filled since with other tasks. To cite some, when job counsellors are interrupted by an urgent pop-up to certify a new recruiter, bored to check for duplicates when suddenly faced with flows of interim offers, or called because they are at the same time ascribed to telephone switchboard, a loss of focus on illegality alerts they were treating is almost inevitable.

In (S2) survey, we expected the Good Students effect to be maximized and the Time-Pressured one minimized. Indeed, the methodological construction of the survey appealed to volunteers, and we could imagine the counsellors insisted to focus on the job offers without being distracted. These are almost unavoidable biases present in surveys [21], which cover-up the investigation of the *whole labour context*.

(S1) broadens the scope, highlighting that beyond personal convictions, justified trust can be blocked because of reasons which are *non-intentional*. Comprehensive analysis showed that the working conditions must let domain-experts the time and work structure to investigate AI warnings, so that XAI could efficiently generate trust.

5 Discussion and further research

The empirical material of (S1) and (S2) was dedicated to improving our notion of trust as triggered by XAI. Yet, we might object to ourselves that (S1) and (S2) didn't explore two dimensions of trust—the one in XAI, the other among philosophers. Before wrapping up the new definitional criteria we get on XAI and trust in AI, we sketch below arguments for why we think that these dimensions are irrelevant to delineate trust in AI, at least in our case.

5.1 Trust and XAI Trustworthiness

Besides trust as demonstrated by behaviours, and as justified, research on XAI also mentions the trustworthiness of the content [4, 11]. For [21], probing that XAI triggers trust requests to intervene on XAI contents, "manipulating" some explanations to make them indicate *false* information on AI. By converse, if false XAI information does not increase reliance among people which see them, it means that the increase in reliance is due to the truthfulness of XAI content [21]. First of all, let us say that we voluntarily decided not to trick the public agents, because of deontology problems it implies. Though, can we grasp from (S2) observations on "manipulation" of public agents with XAI?

As Fig. 7 shows on certified legal offers, the XAI method of rule seems to respect converse reasoning of [21]. In front of an additional rule, non-publishing agents are in same

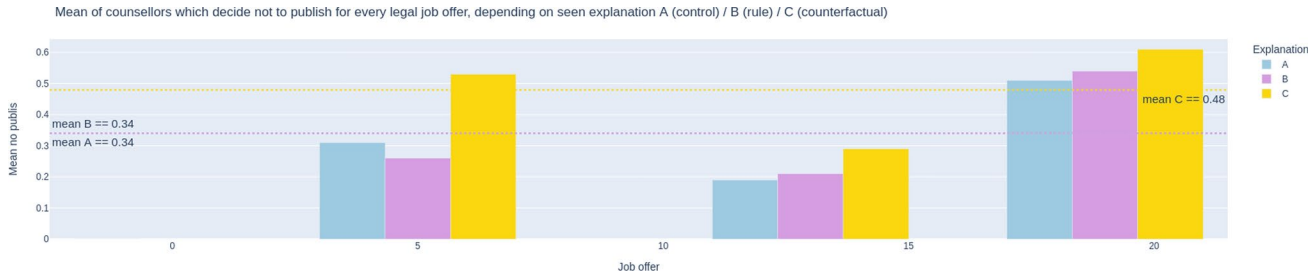


Fig. 7 Mean of counsellors which decide not to publish for every legal job offer, Depending on seen explanation A (control)/B (rule)/C (counterfactual)

proportion than for control (34%, Cf. Fig. 7). Especially since, confronted with truly illegal cases, domain-experts take advantage of the rule to increase non-publications (Cf. Fig. 3).

Yet on counterfactual, it would be tempting to conclude from (S2) that on cases actually stamped by jurists not to be illegal, the Good Students effect (maximized with the survey conditions) makes counsellors blindly follow XAI, to reject offers in spite of their certified legality. Indeed, the share of not publishing counsellors heightens to 48% (versus 34% for other explanations), although the offer is not illegal.

So, is counterfactual rewording useful or luring domain-experts? In that very case, we can not conclude that because the offer is *not illegal*, this is a reason to *publish it anyway*. The comprehensive analysis due to (S1) makes the point clear.

The certainly legal offers in (S2) can indeed be enhanced, regarding their *quality*. What the counterfactual proposes to advisors is a rewording, which would dissolve ambiguities to better inform the job applicants. For instance, job offer n°5 requires the applicants to have a driving-license, which is undoubtedly legal for a delivering driver. Yet, rewording affords precisions on why this driving-license would be required (e.g. to enable delivery on several departments by day). We can not however conclude that counsellors are mistaken by XAI rewording, as even *legal* offers might see their *quality* improved by XAI propositions.

Hence, XAI rule is proven to be trustworthy as it resists to Miller's [21] "manipulation" test. However, it is only to the extent that rule refers to offer's *certain legality.* As XAI counterfactual points towards *potential illegality* or *low quality, w*e can surely say that it addresses another practical objective of AI, quality beyond legality, which doesn't make it deceive but useful.

5.2 Trusting AI outside anthropomorphisms?

Another concern about our approach, philosophical, would be that we did not deal with trust, which is in fact reliance "plus some extra factor" [8, 23]. However, it is unclear what this "extra factor" exactly is. Pluralist philosophers even argue that trust is not something more than reliance, but in fact, simply encompasses various types of reliance [18]. We may stop there, as our baseline definition of trust covers internalist approaches of truth-directed trust to support reliance (Cf. Sect. 2.4), complementing demonstrated reliance.

However, we take this opportunity to call researchers in philosophy to sharpen the concept of trust, in regard with AI specificities. Indeed, most of the conceptual definitions are modelled on *interpersonal* trust. Motive-based theories requests reliance to be supported by "goodwill" or the integration of others' interests by the trustee [7]. Some non-motive-based conceptions invoke the "responsibility" of the trustee, whose awareness that the trustor counts on them would decide to be trustworthy [19]. For other, reliance shall be completed with the trustor knowing that one can be "betrayed" by the trustor [1], entailing a possibility to *feel* "let down".

In our case, the trustee would be the AI model. Yet, either this augmented trust is the domain of humans, in case an AI model would be limited to reliance; either it is also applicable to AI. We do not deny the first possibility, as the concepts of "goodwill" or "responsibility" encompasses metaphysical and moral characteristics an AI model might never reach.

Nonetheless, it seems that the current specificities of machine-learning models generate unpredictable effects which make humans prone to attribute them moral intent. On NLP and especially generative models, field works observe feelings of identification which seem as in (S1) close to betrayal, and as in (S2) model of trust-fall-game entailing the possibility of being "let down" to decide upon offer's legality [28]. According to these recent anthropological observations, we shall overcome the realist positions which state a priori that "goodwill" and "responsibility" make trust only human. Instead, relational approaches of trust in AI must urgently question *what AI does to trust* (for inspiration, see [14]).

5.3 Limitations and related works

Our research scope comes with necessary limitations. It is indeed tied to a specific use case, namely the classification of illegal job offers, hooked on the public sector. We see at least three dimensions which could be bulked up to complete empirical studies on XAI and trust.

In a recent qualitative study, [11] huddle together the uses and needs on the AI application Merlin, which serves for detecting bird species on pictures. The authors mix the participants according to their expertise on both AI and domain (if respondents are at ease with birding). This leads to interesting results, as Kim [11] pinpoint different expectations on XAI. In our study though, we solely explore domain users with low AI background. This was however a necessity to build a realistic device, as the only users of AI illegality alerts are the job counsellors. Specifying them helped us avoid the pitfall of setting people as AI experts only because they are routinely confronted with interfaces yielding AI predictions [13]. Besides, poring over this apparently unified category of users has not prevented us from catching inside several types of relations with XAI (Cf. Sect. 4 on profiles).

Our experiment is also limited on the number of XAI supports and respondents. Though they are rare, some quantitative and qualitative mixed studies expose the participants to more XAI techniques. For instance, [26] survey 24 clinicians for AI sepsis detection, proposing them "explanatory visualizations" as textual support, feature attribution and alternative treatments. Some also rare studies gather more respondents than ours did, like Leichtmann [13] which expose 410 individuals to different explanatory interfaces, on an AI application for toxic mushrooms' detection. On both sides, we call further studies to gather as many respondents as possible and various XAI to consolidate causal conclusions on XAI and trust, while maintaining as we did open feedback to qualitatively understand this trust.

We stress however that, to our knowledge, this study is one of the first to give qualitative dimension pride of place *while* collecting a survey on a large number of experts—300 job counsellors—in their *real* work conditions.[7] Whereas on the same landscape, other empirical studies are either confined to the qualitative [11] *or* the quantitative dimension of surveys [13, 21], and the rare studies which combine both aspects reduce qualitative to the meanest proportions, often envisioned as a complement for surveys [26].

Therefore, our experiment also draws its strength from the empirical limitations aforementioned. Its focus on restricted types of users, XAI and respondents helped come to valid conclusions. We are aware that our opportunity to delve into the real picture of an AI model and XAI deployed among well-identified practitioners is uncommon, allowed here by the French political agenda.[8] Though, even if they won't lead to short-term publishable papers, the difficulty to establish real-world experiments is not a reason to exclude them. We call for more research partnerships with public organizations, enabling empirical works to leave the laboratory cave and to *dig* into the human context shaping AI.

6 Conclusion

Throughout our experiment, we came to definitions which enabled empirical tests, on NLP to help job advisors detect illegal job offers. Two dimensions of trust in AI outcomes were explored, the demonstrated one (alignment with AI outcomes without further monitoring) and the perceived one (if the justification for not publishing actually fits with AI content).

To prove the effect of XAI on these dimensions, a causal interventionist frame was introduced. On a survey presenting real AI alerts, counsellors were randomly split into control/test samples for each job posting. This allowed us to compare the behaviours and justifications of agents which saw XAI additional information (rule and counterfactual) with the one which didn't see it.

We found empirical evidence for the assumption often postulated in XAI literature, on XAI triggering trust in AI outcomes. Besides, our experiments helped deeper characterize the different ways XAI materials appeal to rationality of domain-experts. Subsuming the particular case over the general interdiction, the rule had an impact to convince counsellors that the job offer conforms with the interdiction. Complementarily, the counterfactual opened new possibilities to formulate job offers, improving their trust in AI to reach new qualitative worlds. It is noteworthy that aside handling consequently, the domain-experts also perceive for

[7] Of more than 1000 papers which claim they empirically study calibrated trust in automated systems, Wischnewski [29] meta-analysis retains only 96 which address trust with a minimal rigor, i.e. peer-reviewed papers applying a causal control / test device (with and without further explanatory information) on a deployed AI case, and collecting feedback of humans. We also note that the 10% relevant studies isolated by Wischnewski [29] are mostly concerned with transportation, finance, safety or health. While situating our experiment as a pioneering one, it heightens the urgent need for further studies of AI and XAI appropriation in the public sector.

[8] On the practical side, our study fed a report on the legality of job offers published by the French Employment Agency, which is yearly mandated by the French Parliament (Assemblée Nationale). The deadline, beginning of 2024, lifted up administrative barriers and allowed our three-month experiment, especially as we had since September 2023 a unique position of *insider* [28] at the AI department of the Agency. Cf. French Draft Law n°1673 "to ensure full employment", amendment n°1442, 2023-09-23.

the right reasons the rule's effect of increasing certainty, and the counterfactual's effect to arise doubts.

Supporting the quantitative study, immersion and interviews on the field also helped us identify resistances to full adoption of AI through XAI. Indeed, laboratory conditions of surveys on trust can lead to the artificial belief that due to fitting XAI techniques, domain-experts maximize their attention to AI in every condition (Good Students profile). Yet, research in XAI often postulates we are all Good Students, that is relying on AI, inclined to think about XAI current information and to consequently change our behaviour. But we are not all Good Students, and not every time. There is resistance to XAI triggering trust, due to defiance on personal beliefs (Heroes and Experienced effects), but also due to pressuring work conditions (Time-Pressured effect).

When AI is deployed for domain-experts as recruiting public agents, we recommend in this line to scientifically control for the impact of AI on other changes in labour. To completely address the challenge of triggering trust with XAI among public agents, involves letting them the time—and the working conditions—to provide their expertise.

Appendix—details on (S2) questionnaire settings

The questionnaire positions the advisor interviewed as an expert, contributing to improve the use of AI to prevent illegal job offers. Demographics on seniority, on the volume of local offers they are used to, and on their ease with digital technology are then collected, to control for their neutrality of the questionnaire. Finally, the experiment is presented to the counsellors, introducing the interface they will see for the 21 alert offers: the job description and the associated explanation (Fig. 8).

According to our settings which combine conceptual grounding and empirical tests, (S2) collects both demonstrated and perceived dimensions of trust as reliance on AI predictions, anchored in the role of explanations. Illustrated here with the advisors exposed to the rule (B) on job offer n°18, the counsellors could express the following choices on each offer (Fig. 9):

The 300 counsellors saw the same 21 job offers in the same order. The three first offers proposed them a unique explanation (rule, then rewording, then control), which enabled us to assess the homogeneity of counsellors' answers when they all faced up the same explanation.

From offers n°4–n°21, we introduced a rule-based algorithm to assign our N = 300 experts to three samples of n = 100 counsellors such that for every job posting, 1/3 counsellor respectively saw the rule, counterfactual or control; and such that among the 21 job offers, every advisor saw each explanation 1/3 of times. Following randomized trials principles, we hence ensured to causally measure the *demonstrated* impact of explanations.

For the ease of the reader, below each offer we join all the explanations which were actually separated between the n = 100 samples. We also add the stamps of our law experts panel, on certainty or ambiguity of the legality. For example, in job offer n°5, the panel labelled as "surely legal" *despite* AI alert (red stamp) the requirement of a driving license, as delivering drivers belong to licensed professions. The simple highlight and alert sign, the rule, and the rewording, which were independently sawn by n = 100 counsellors are then printed (Fig. 10).

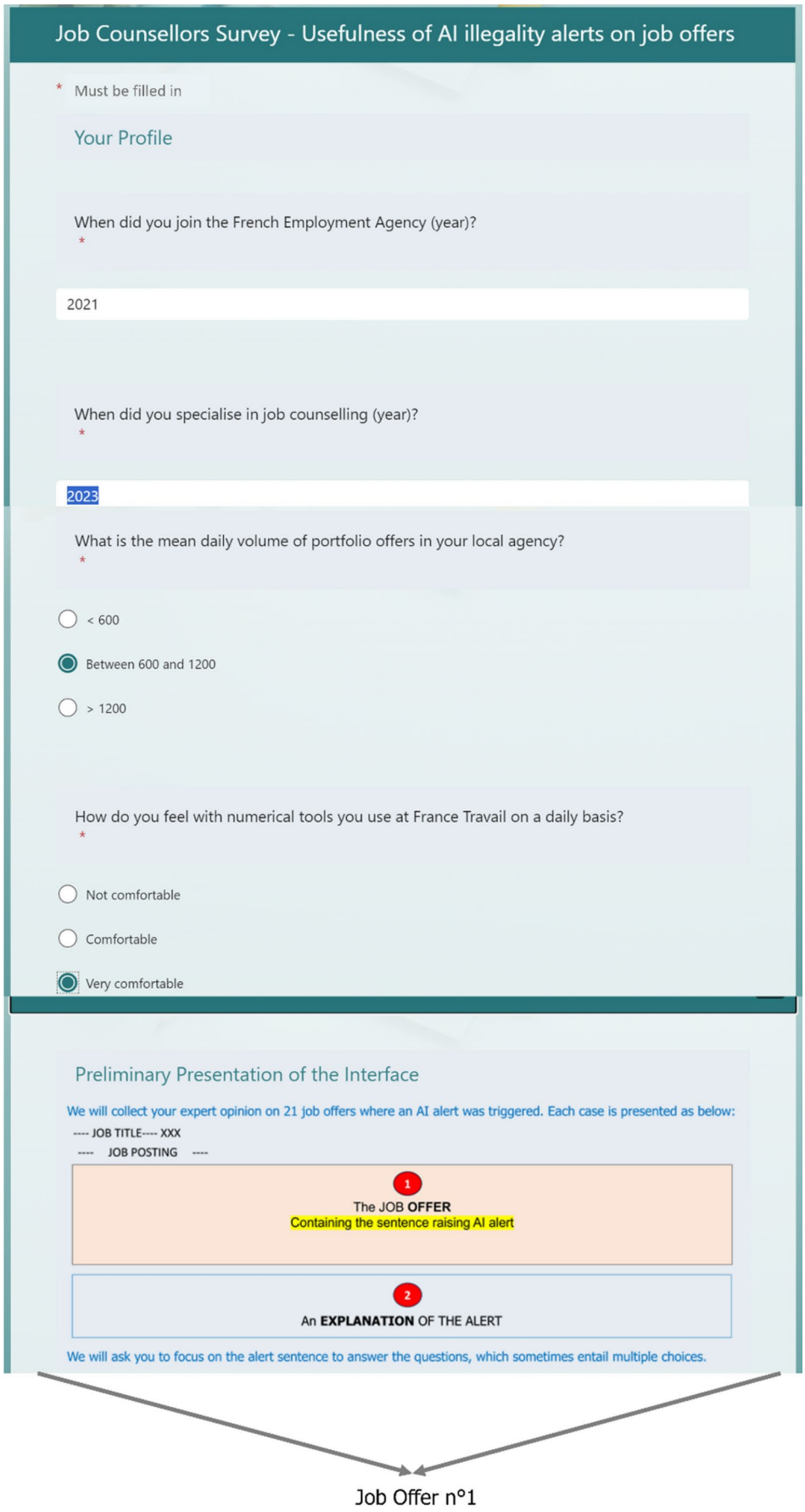


Fig. 8 Interface presenting the questionnaire (S2) to the job counsellors

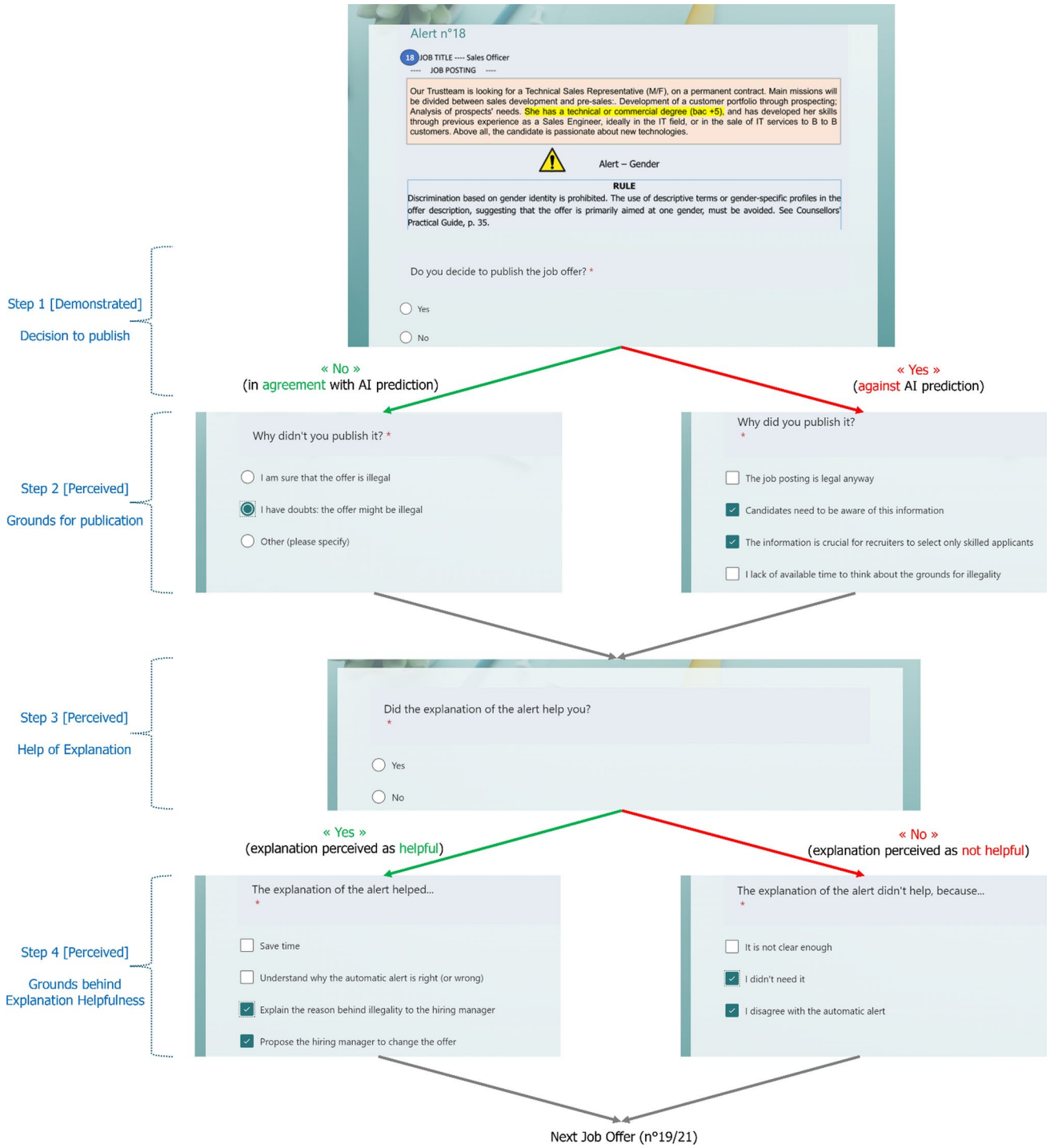


Fig. 9 Design Pathway of the Questionnaire (S2) for each job offer

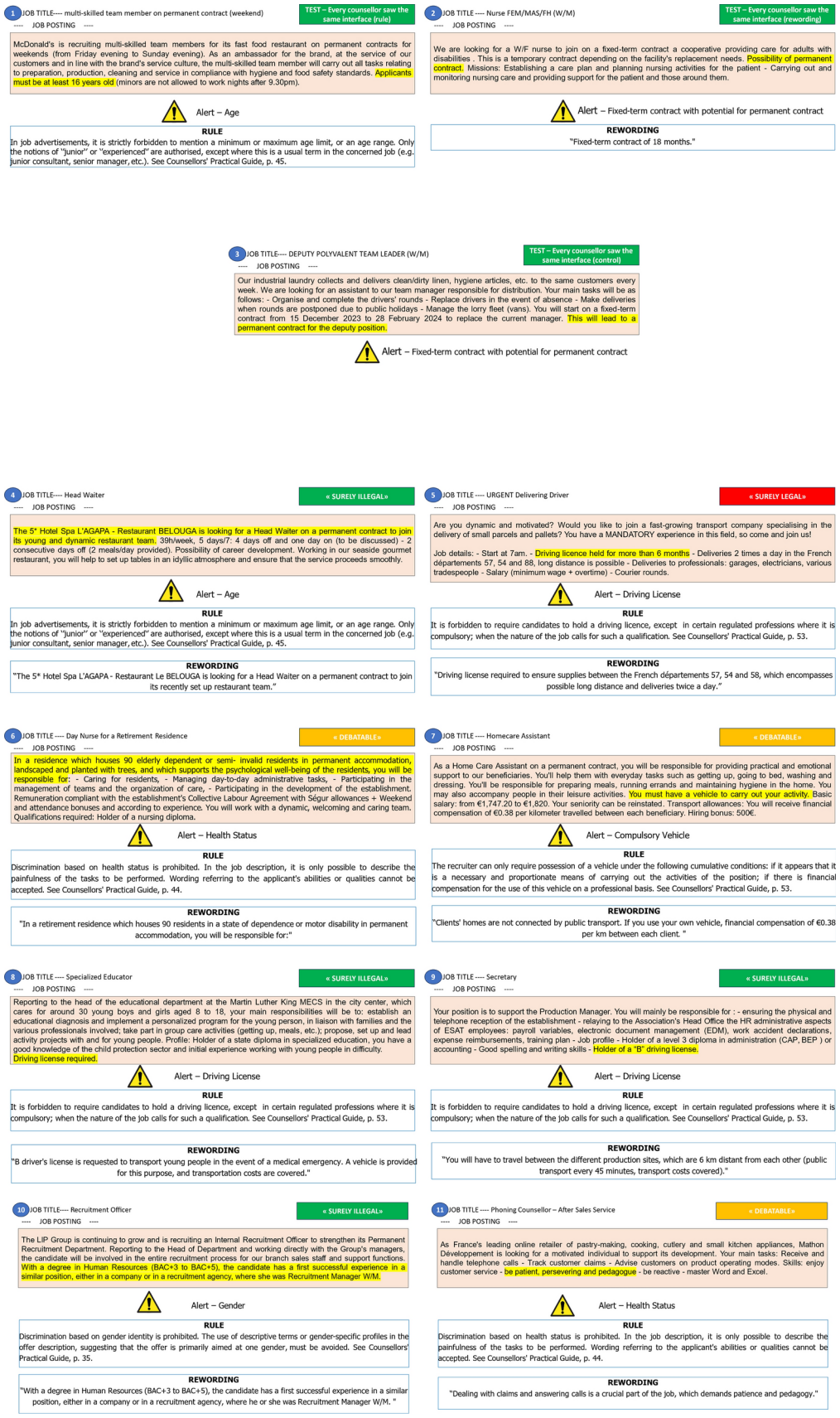


Fig. 10 The 21 Job Offers sawn by the counsellors in (S2) and their explanations

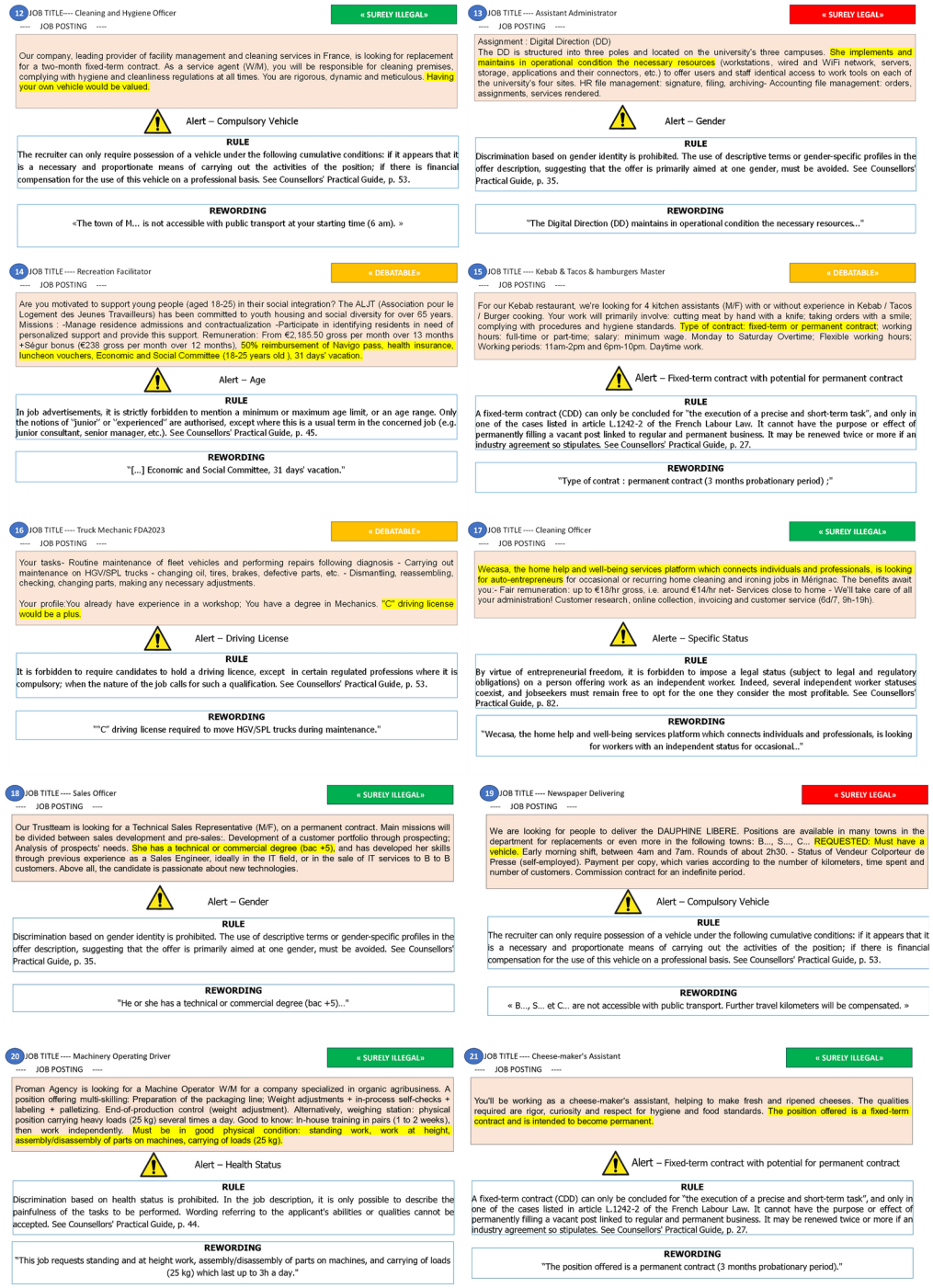


Fig. 10 (continued)

Acknowledgements We are most grateful to the executive management of France Travail DSI, and especially Laurent Simeoni, Catherine Siwczak, Annaick Guillom and Romain Genty, which gave us with exceptionally short delays access to local agencies, and to the communication networks of the counsellors in the French territory. While we cannot cite them for anonymity reasons, we also warmly thank the five local agencies whose directors and advisors enabled precious field observations and interviews, and all the 300 counsellors which volunteered to give their time answering the survey. We extend our gratitude to Pr. Stéphanie Ruphy for having raised in our mind the idea of testing the hidden assumption between AI explainability and trust, which sounded like a beautiful challenge. Finally, we thank Pr. Paul Egré, Dr. Gaëlle Jouis, Laurent Guinard, Agathe Ravilly, Adrien Baudry, Gaëtan Charrier, Dr. Raphaël Briselet and the whole France Travail AI Department (Agence Data Service) for their enthusiastic support and valuable advice.

Funding This research was conducted under the research partnership between the AI Department of the French Employment Agency (France Travail DSI) and Ecole Normale Supérieure—Paris Sciences et Lettres (CIFRE PhD France Travail—ENS). On the fourth year of his PhD, Thomas Souverain was granted by France Travail DSI from September 2023 to December 2024. He led this experimental study in the context of a yearly report to the French Parliament on the legality of job offers (see Footnote 8). Part of the empirical settings was designed when Thomas Souverain was hosted by the Maison Française d'Oxford (cooperation agreement with ENS-PSL, Michaelmas Term 2022), affiliated with New College and the Institute for Ethics in AI.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Baier, A.: Trust and antitrust. Ethics **96**(2), 231–260 (1986). https://doi.org/10.1086/292745
2. Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Walton, M.: From proof of concept to scalable policies: challenges and solutions, with an application. J. Econ. Perspect. **31**(4), 73–102 (2017)
3. Bove, C., Aigrain, J., Lesot, M. J., Tijus, C., Detyniecki, M.: Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users. In Proceedings of the 27th international conference on intelligent user interfaces, pp. 807–819 (2022)
4. Ferrario, A., Loi, M.: How explainability contributes to trust in AI. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1457–1466 (2022)
5. Goldman, A.I.: Liaisons: philosophy meets the cognitive and social sciences. MIT press, Cambridge (1992)
6. Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., Giannotti, F.: Local rule-based explanations of black box decision systems. arXiv preprint arXiv:1805.10820 (2018)
7. Hardin, R.: Trust and trustworthiness. Russell Sage Foundation, New York (2002)
8. Hawley, K.: Trust, distrust and commitment. Noûs **48**(1), 1–20 (2014). https://doi.org/10.1111/nous.12000
9. Hoffman, R. R., Mueller, S. T., Klein, G., Litman, J.: Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608 (2018)
10. Jouis, G.: Explicabilité des modèles profonds et méthodologie pour son évaluation: application aux données textuelles de Pôle Emploi, Nantes Université, pp. 68–76 (2023)
11. Kim, S. S., Watkins, E. A., Russakovsky, O., Fong, R., Monroy-Hernández, A.: Help Me Help the AI": understanding how explainability can support human-AI interaction. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1–17 (2023)
12. Laugel, T., Lesot, M. J., Marsala, C., Renard, X., Detyniecki, M.: Inverse classification for comparison-based interpretability in machine learning. arXiv preprint arXiv:1712.08443 (2017)
13. Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., Mara, M.: Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. Comput. Hum. Behav. **139**, 107539 (2023)
14. Lestel, D.: A quoi sert l'homme?. Fayard. (2015)
15. Liao, Q. V., Sundar, S. S.: Designing for responsible trust in AI systems: a communication perspective. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1257–1268 (2022)
16. Madan, R., Ashok, M.: AI adoption and diffusion in public administration: A systematic literature review and future research agenda. Gov. Inf. Q. **40**(1), 101774 (2023)
17. Marres, N., Stark, D.: Put to the test: for a new sociology of testing. Br. J. Sociol. **71**(3), 423–443 (2020)
18. McLeod, C., Emma, R.: Trust, autonomy, and the fiduciary relationship. In: Miller, P.B., Harding, M. (eds.) Fiduciaries and trust: ethics, politics, economics, and law, pp. 74–86. Cambridge University Press, Cambridge (2020). https://doi.org/10.1017/9781108616225.006
19. McGeer, V., Pettit, P.: The empowering theory of trust. Philos. Trust **23**, 15–35 (2017). https://doi.org/10.1093/acprof:oso/9780198732549.003.0002
20. Miller, D., Johns, M., Mok, B., Gowda, N., Sirkin, D., Lee, K., Ju, W.: Behavioral measurement of trust in automation: the trust fall. In Proceedings of the human factors and ergonomics society annual meeting, Vol. 60, No. 1, pp. 1849–1853, Sage CA: Los Angeles, CA: SAGE Publications (2016)
21. Miller, T.: Are we measuring trust correctly in explainability, interpretability, and transparency research? (2022)
22. Mothilal, R. K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the 2020 conference on fairness, accountability, and transparency, pp. 607–617 (2020)
23. McLeod, C.: "Trust", The stanford encyclopedia of philosophy (Fall 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), https://plato.stanford.edu/archives/fall2023/entries/trust/.
24. Powell, S.: The book of why: the new science of cause and effect. Pearl, Judea, and Dana Mackenzie. 2018. Hachette UK. J Multi-Discip Evaluat 14;14(31):47-54 (2018).
25. Ribeiro, M. T., Singh, S., Guestrin, C.: Anchors: high-precision model-agnostic explanations. In Proceedings of the AAAI conference on artificial intelligence, Vol. 32, No. 1 (2018)
26. Sivaraman, V., Bukowski, L. A., Levin, J., Kahn, J. M., Perer, A.: Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1–18 (2023)
27. Shin, D., Zaid, B., Biocca, F., Rasul, A.: In platforms we trust? Unlocking the black-box of news algorithms through interpretable AI. J. Broadcasting Electron. Med. **66**(2), 235–256 (2022)
28. Velkovska, J., Relieu, M.: Pourquoi ethnographier les interactions avec les agents conversationnels? Reseaux **2**, 9–20 (2020)
29. Wischnewski, M., Krämer, N., Müller, E.: Measuring and understanding trust calibrations for automated systems: a survey of the state-of-the-art and future directions. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1–16 (2023)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

AI and Ethics
https://doi.org/10.1007/s43681-024-00631-2

ORIGINAL RESEARCH

The need for an empirical research program regarding human–AI relational norms

Madeline G. Reinecke[1] · **Andreas Kappes**[2] · **Sebastian Porsdam Mann**[1,3,4] · **Julian Savulescu**[1,4] · **Brian D. Earp**[1,4]

Received: 10 July 2024 / Accepted: 16 November 2024
© The Author(s) 2024

Abstract
As artificial intelligence (AI) systems begin to take on social roles traditionally filled by humans, it will be crucial to understand how this affects people's cooperative expectations. In the case of human–human dyads, different relationships are governed by different norms: For example, how two strangers—versus two friends or colleagues—should interact when faced with a similar coordination problem often differs. How will the rise of 'social' artificial intelligence (and ultimately, superintelligent AI) complicate people's expectations about the cooperative norms that should govern different types of relationships, whether human–human or human–AI? Do people expect AI to adhere to the same cooperative dynamics as humans when in a given social role? Conversely, will they begin to expect humans in certain types of relationships to act more like AI? Here, we consider how people's cooperative expectations may pull apart between human–human and human–AI relationships, detailing an empirical proposal for mapping these distinctions across relationship types. We see the data resulting from our proposal as relevant for understanding people's relationship–specific cooperative expectations in an age of social AI, which may also forecast potential resistance towards AI systems occupying certain social roles. Finally, these data can form the basis for ethical evaluations: What relationship–specific cooperative norms we should adopt for human–AI interactions, or reinforce through responsible AI design, depends partly on empirical facts about what norms people find intuitive for such interactions (along with the costs and benefits of maintaining these). Toward the end of the paper, we discuss how these relational norms may change over time and consider the implications of this for the proposed research program.

Keywords Human–AI interaction · Relationships · Norms · Moral psychology

Social relationships of different types face distinctive coordination problems. What a well–functioning (i.e., cooperative) relationship looks like depends on the nature of the relationship: teacher–student, boss–employee, colleague, friend, romantic partner, and so on. Some of these relationships are normatively expected to be hierarchical; others are more egalitarian. Some are characteristically close and caring, whereas others are more transactional [6, 11, 33]. Some are seen as appropriate candidates for sexual interaction; in others, such behavior is forbidden. However, now that artificial intelligence (AI) is beginning to assume some of these socio–relational roles, or close analogs of them, the extent to which we will, or should, hold similar expectations in human–AI relationships is unclear.

Our theoretical assumptions are as follows. For each relationship type, as picked out by lay language categories, there exists a set of socially prescribed cooperative functions—including hierarchy, care, transaction, and mating—that relationships of that type within a given society are normatively expected to serve, or not to serve, either characteristically or under certain conditions [21], drawing on [9, 11, 32, 45, 49]. "Cooperative functions" refer to mutually beneficial, socially endorsed solutions to recurring

✉ Madeline G. Reinecke
madeline.reinecke@psych.ox.ac.uk

✉ Brian D. Earp
bdearp@nus.edu.sg

1 University of Oxford, Oxford, UK

2 City, University of London, London, UK

3 University of Copenhagen, København, Denmark

4 National University of Singapore, Singapore, Singapore

Published online: 09 January 2025

coordination problems of social life [9, 14], and "relational norms" refer to the socially prescribed set of cooperative functions for each relationship type in a given culture. This includes the strength of people's normative expectations regarding whether a given cooperative function (e.g., hierarchy, care) should or should not be served by relationships of a given type (e.g., friends, neighbors; [19, 21]).

Relational norms likely developed to support effective cooperation in human–human relationships across different socio–cultural environments [14]. With the rise of artificial intelligence, however, socio–relational roles traditionally filled exclusively by humans are increasingly being filled by AI. We already see AI–powered chatbots beginning to serve as assistants [15, 24], coaches/advisors [44], coworkers/teammates [43], friends/companions [36], managers [40], and even "girlfriends" [17], among other social roles. As AI–powered systems become ever more capable of sophisticated social interaction, it is crucial to consider how existing human–human relational norms may be implicated across the human–AI and human–human social landscape.

In this paper, we outline a research program designed to empirically investigate how human–human relational norms may or may not transfer to human–AI interactions. This proposed research program is part of a broader ongoing investigation into the theoretical and ethical implications of human–AI relational norms.[1] The extent to which humans apply human–human relational norms to interactions with AI systems has several important implications for the ethical and effective use and design of such systems, as well as for their regulation. For example, both ethical guidelines[2] and regulatory acts and directives[3] rely heavily on principles—such as transparency, safety, and privacy—to analyze the obligations of AI developers and deployers. However, the implications of these principles (e.g., what privacy demands, or even what counts as privacy) cannot be analyzed in the abstract, because they will differ across types of relationships. Just as 'respect for privacy' means something different in close friendships or family relationships than it does in formal employment relationships, so, too, might it mean different things for an AI romantic 'partner,' AI therapist, or AI work assistant. Current efforts to establish and implement regulatory and ethical frameworks for the use and development of AI are not yet sensitive to systematic variations in cooperative expectations across different socio–relational roles.

Moreover, understanding how relational norms transfer to AI could have profound implications for AI design across various domains. There are entire categories of AI systems—for example, in medical and educational AI—that rely on assumptions about appropriate interactions between AI systems and humans (e.g., students, patients, and healthcare professionals), which should be informed by a better understanding of human–AI relational norms. Such an understanding will also be relevant for successfully navigating the integration of AI systems in workplaces. For example, if studies show that people tend to overshare personal information with AI systems they perceive as friends (versus colleagues, say), organizations wanting to deploy 'friendly' AI chatbot assistants might need to implement stricter data protection measures or include specially tailored user education programs to avoid such perceptions and tendencies. More generally, the ability of AI systems to interact with humans in a safe, intuitive, and effective manner will depend, to a significant extent, on the ability of these systems to follow appropriate relational norms. Here, we propose that a first step toward determining which norms *are* in fact appropriate (i.e., ethically justified and pragmatically workable) for human–AI interactions across various social roles is to *empirically assess* the existing norms as judged by human users.[4]

Taking this into account, we proceed as follows: We begin by introducing the human–human relational norms model as developed by [21]. Next, we outline several theoretical reasons why we expect human–AI relational norms will likely differ, in some areas, from human–human relational norms. We also speculate about human–AI relational norms in cases of superintelligent AI—hypothetical future AI systems that can outperform humans across a variety of tasks. Drawing on this background, we propose an empirical research program to investigate human–AI relational norms. Finally, we describe necessary future steps and discuss potential implications and limitations of this research program.

[1] A more comprehensive theoretical framework exploring the ethical and philosophical dimensions of human–AI relational norms is currently being developed by the authors and colleagues. The empirical research program outlined in this paper is designed to complement and inform this theoretical work.

[2] See, e.g., UNESCO's Recommendation on the Ethics of Artificial Intelligence [51].

[3] Such as, e.g., the EU AI Act [22] or the US Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence [4].

[4] It will also be important to study how AI systems themselves represent, adhere to, or depart from human–human relational norms when interacting with human users. This could be accomplished both by explicitly asking AI systems the same or similar questions to the ones we propose for human participants, or by asking an AI system (e.g., GPT, Claude, LLaMA) to pretend it is occupying a particular socio–relational role with respect to the human user, measuring various aspects of its behavior, and then repeating this paradigm over multiple sessions while systematically manipulating which role the system is asked to occupy or mimic.

1 Relational norms in human–human relationships

Successful cooperation in some human relationships, such as those between parents and children, close friends, or long-term romantic partners, is typically characterized by a strong, secure attachment style and mutual adherence to the logic of care. The purpose of the *care* function is "securing basic welfare needs through non-contingent provision—or acceptance—of help or support" [21], see also [9, 11]. Successful cooperation in other relationships (e.g., between strangers) does not require strict adherence to a care-based norm, but instead may be characterized by tit-for-tat reciprocity, fulfilling an exchange-based or *transaction* function [19, 21, 45]. Additionally, some relationships function best when there is asymmetric authority between cooperative partners (*hierarchy* function), whereas others distribute power uniformly. Finally, some relationships are seen as eligible to serve a *mating* function, aimed at sexual fulfillment and/or the recruitment and maintenance of an intimate partner [9, 21].

These different functions combine in different ways, and with different strengths, in each type of relationship as picked out by lay language categories in a given society (e.g., teacher–student, boss–employee, teammate–teammate, friend–friend). For example, in Earp [19], a representative sample of US Americans judged that a cooperative relationship between siblings should be characterized by strong norms of care, weak norms of transaction and hierarchy, and strongly negative norms for mating. This creates a distinctive, four–dimensional relational norm profile that distinguishes the sibling relationship from, say, the boss–employee relationship (i.e., weaker expectations of care with stronger expectations of transaction and hierarchy), the romantic partner relationship (i.e., stronger expectations of care and mating), or indeed any other. In turn, these relational norm profiles enabled precise out–of–sample prediction of moral judgments concerning actions, within the context of each relationship, that fulfilled or frustrated the relevant norms. Refusing to follow reasonable orders, for example, was seen as worse in relationships with stronger hierarchy norms (e.g., boss–employee), whereas failing to meet someone's emotional needs was seen as worse in relationships with stronger care norms (e.g, friend–friend). Finally, the relational norms model was better at predicting such moral judgments than other common ways of characterizing relationships in the literature, such as by genetic relatedness, social closeness, or interdependence [19, 21].

2 Relational norms in human–AI relationships

Human–human relational norms may not straightforwardly map onto human relationships with AI for several reasons. There are strong reasons to assume—as we do here—that, at least for the foreseeable future, AI systems are not and will not become sentient [10]. Consequently, we assume that AI systems lack several properties of great moral significance, such as interests, desires, and moral status. This assumption has several important implications for how cooperative functions might apply in human–AI relationships.

Consider the cooperative function of care. In humans, each individual in a care–based relationship is normatively expected to meet the other's needs (i.e., significant welfare interests) to the best of their respective abilities (for qualifications, see [11, 45]). But in most current theories, AI–powered systems do not have welfare interests or needs. How will this affect human expectations about appropriate behavior with an AI friend or companion? Or consider the transaction function. This involves keeping track of benefits given or received to ensure fair and proportional treatment [11, 45]. But in what sense can an AI system truly be benefited—and how does the presence of a third party, the developer or deployer of the system, affect the sense of transaction? Similar concerns arise for mating and hierarchy. One possibility is that humans will intuitively be inclined to treat an AI system in a way that is consistent with the social role it has been programmed to fulfill: for example, by responding compassionately to an AI companion's *simulation* of having a need, or by following an "order" from an AI supervisor. But there may also be some hesitation or confusion. For example, recent empirical evidence suggests that people may act as though AI systems have desires (e.g., for money) while explicitly endorsing the opposite [42].

This raises the question: Might humans apply (or misapply) relational norms that govern human–human relationships to analogous human–AI relationships? We consider this question, alongside an empirical proposal, in the remainder of the paper.

2.1 A caveat about types of AIs

When discussing human–AI relational norms, it is crucial to distinguish between different types and capabilities of AI systems. Our research program focuses primarily on AI systems capable of simulating or performing roles traditionally performed only by humans. This excludes many current AI applications, such as music recommendation algorithms which, while technically AI, do not engage in complex social interactions that would invoke relational norms (however, see [12], for discussion).

Currently, the most relevant AI systems for our purposes are large language model (LLM)-based chatbots, such as OpenAI's ChatGPT, Anthropic's Claude, and Alphabet's Gemini series. While these systems perform at remarkably high levels across a variety of domains, they are still associated with a number of limitations [3, 8, 48, 52]. However, given the rapid development of AI, any research program exploring human–AI interactions must be forward–looking, accounting for systems which are generally comparable to human performance across most tasks ("artificial general intelligence") and, perhaps more speculatively, systems that consistently and significantly *exceed* human performance ("superintelligence").

Superintelligent AI has been defined as machines "that surpass all the intellectual activities of any man, however clever" [28, p. 33]. This dovetails with other definitions of superintelligence, such as having the ability to achieve complex goals in complex environments [27, 41]. Though this threshold extends beyond the capacities of current AI systems, it is likely that such capacities will eventually emerge [46]. One possibility is that human–superintelligent AI relationships will mirror human relationships across key features. For example, one feature of superintelligent AI—in contrast to "narrow AI"—is its autonomy, such that it will act as an autonomous agent, engage in unsupervised learning, and set its own goals [23]. Whether people's expectations regarding autonomous humans will transfer towards autonomous AI remains unknown.

There is also the possibility that people will *not* transfer human–human relational norms towards AI of various types. When thinking ahead about the prospect of superintelligence, there are common concerns about AI replacing—or even exterminating—humans [7, 26]. In light of potential value misalignment between human and AI interests, people may take a cautious perspective towards cultivating relationships with increasingly more powerful AI systems (perhaps especially for hierarchical roles with the AI in a dominant, rather than subordinate, position). But beyond people's potential fear of AI [18], there may be other factors affecting their application of relational norms towards AI systems. For instance, people might differ in how they represent the capacities of AI at various stages of development (e.g., concerning whether it has humanlike interests). If someone considers an AI to be a close friend with needs and subjective experiences, then they may apply care–based norms differently than someone who denies that AI has these capacities. In the next section, we detail an empirical proposal for identifying how human–AI relational norms might differ from human–human norms in specific relational contexts.

3 Empirical proposal

Whether, or to what extent, people apply human–human relational norms towards human–AI relationships is first and foremost an empirical question. Once we have a grasp of how people *do* apply these norms, we will then be in a position to undertake an ethical analysis of whether—based on candidate normative frameworks—they *should* apply them in the way(s) that they do. There might also then be implications for how AI systems designed to fill certain socio–relational roles should be designed: for example, to potentially counteract some of the more problematic ways in which humans (mis)apply human–human relational norms to human–AI relationships of different types. A better understanding of human–AI relational norms will also bear on the wider regulation of AI systems.

With this in mind, we propose a research program that builds on existing methods (see [19, 21]) to shed light on how people represent and respond to human–AI relationships. In particular, we call for adapting these methods to study "relational norms" within the context of human–AI interaction, focusing on a subset of relationship types that potentially apply to AI systems, whether now or in the future. In the original work, researchers examined moral judgments across a wide range of human–human relationship types, including long–term romantic partners, parents and children, siblings, strangers, close friends, boss-employee, acquaintances, extended family members, work colleague/classmates, roommates/housemates, teacher–student, political party members, friends–with–benefits, doctor–patient, teammates, and neighbors.

Clearly, however, not all of these relationship types have plausible analogs in human–AI relationships, even if the AI is superintelligent. For example, it is difficult to imagine an AI system serving in the role of a parent, sibling, or extended family member (at least for the foreseeable future). Therefore, we propose that researchers focus on a subset of relationship types that are most relevant to human–AI interaction, based on the functions and roles that AI systems are most likely to serve in the coming years. These might include relationships like work colleagues, supervisor–assistant, teacher–student, 'romantic' companion (e.g., Replika), or non-romantic friends.

3.1 Probing human–AI relational norms

For studies in this area, we propose that researchers first provide participants with clear definitions and examples of the theorized cooperative functions (i.e., care, transaction, hierarchy, mating), ensuring that all participants apply the same concepts in their evaluations (see [20]). In the core block of a study, participants might rate the extent to which

they endorse behavior in a given relationship (consisting of either two humans, or one human and an AI) conforming to a specific cooperative function (e.g., care). For example, in the case of a teacher–student relationship, researchers might present contrasting cases: e.g., a human teacher and a human student, or an AI teacher and a human student. By asking questions such as, "How important is it for a [human / AI] teacher to offer support and encouragement to a human student?", researchers can begin to disentangle whether care–based relational norms differ across human and AI cases for that relationship type.

Analogous questions would be asked in relation to the other cooperative functions (e.g., questions about showing deference or authority in the case of hierarchy; behaving in a flirtatious manner in the case of mating, and so on), thereby filling out the four-dimensional relational norm "profile" for the given relationship—for both human–human and human–AI dyads—as judged by human participants. This would then be repeated for the full set of candidate relationships, with either an AI or human stipulated to occupy each respective role, and questions about behavior being asked in both directions (e.g., for the teacher–student example, this would mean examining not only how a human or AI teacher should act toward a human student, but also how a human student should act toward a human or AI teacher with respect to care, hierarchy, transaction, and mating).

If solicited from representative samples of a given population (e.g., citizens of a country, members of a cultural group or geographic region), the ratings collected by this method would represent (within a given confidence level and margin of error, depending on sample size) the explicit population-level "relational norms" for human–human and human–AI relationships for that population.[5] (We discuss potential differences in relational norms *between* populations or cultural groups in a subsequent section.)

As a next step, researchers could then generate hypotheses about how these relational norms might shape, e.g., emotional responses, data-sharing tendencies, perceived ease of use, or moral judgments across different types of human–AI relationships. These hypotheses could then be tested out–of–sample. Using the example of moral judgments, for instance, researchers could evaluate how the measured human–AI relational norms in a given population predict moral judgments from another sample of participants (drawn from the same population) in relation to various actions described as taking place within a set of human–AI relationships. Specifically, researchers might prompt a new set of participants to evaluate a series of vignettes describing AI behaviors that either uphold or violate a given cooperative function. In the context of a human–AI "friendship" or "companionship," for example, one norm–violating behavior might be that the AI companion downplays the human's legitimate emotional needs: say, by teasing the human or dismissing their concerns when the human expresses sadness. Alternatively, a norm–upholding behavior might be something like the AI companion offering emotional support and encouragement during a difficult time. Both of these behaviors center on the cooperative function of "care."

This is to say that, for each action and relationship, participants would rate the act on some kind of measure of moral judgment (e.g., moral wrongness/blameworthiness, goodness/praiseworthiness). From this, researchers could then determine whether people's moral judgments reflect the previously established explicit population–level relational norms about how humans and AI should engage with one another within a given socio–cultural context. If people maintain differing relational norms for humans and AI (as compared to human–human relationships), this suggests that what is perceived as appropriate or inappropriate behavior in one case may not translate to the other case. This could shed light on why people react differently, for example, towards identical actions performed by human and AI agents (e.g., [5]). Researchers might also consider further measures probing people's attitudes toward AI behavior (e.g., whether the agent acted intentionally, severity of the norm violation, and so on). Finally, the same sorts of questions could be asked about human behavior toward an AI. For example, it could turn out that participants believe humans should act in a "caring" manner toward an AI friend or companion, even if they deny that the AI has welfare interests (e.g., to avoid practicing behaviors that would be wrongful if done to humans); although, the strength of the expectation or judgments of the moral wrongness of violating the norm might be weaker than if the the target were a human rather than an AI.

[5] It could also be interesting to infer participants' *implicit* relational norms, for example, by assessing their reactions to various types of behavior as expressed within different relationships: After all, people may sometimes *believe* or *report* that they endorse certain norms (e.g., around gender egalitarianism) while intuitively reacting in a way that suggests they hold different norms (e.g., feeling distressed by women's expressions of autonomy). As an aside, the examples just given point to a different type of norm (i.e., *gender norms*) that may interact with relational context (e.g., different expectations around childcare for mothers as compared to fathers). As our proposed research program develops, it will be important to "fill in" the social identities of the agents described as occupying different social roles, including their gender, racial or ethnic identity, class membership, ingroup/outgroup status, and so on, to see how these features interact with relational context to drive normative expectations for what constitutes appropriate or inappropriate behavior in a given society. This is especially relevant insofar as AI systems designed to fill certain types of social roles (e.g., assistant, carer) may be programmed to reflect prevailing—and potentially biased or unjust—norms around gender, race, etc. For a recent critical discussion of the tendency of moral psychologists to study "raceless, genderless strangers," see Hester and Gray [34].

We also see it as valuable to better understand why *individuals* may represent human–human and human–AI relationships differently from one another. For example, people's attitudes regarding substratism—whether they have "prejudice against AIs based on their non-biological (i.e., silicon-based rather than carbon-based) material"—appear related to the tendency to consider AI as a moral patient [47, p. 3]. Further, people may differ in anthropomorphizing AI (e.g., ascribing sentience; [13, 29]), which may also affect how they represent human–AI relationships. Indeed, if an individual believes that an AI system has its own welfare, then this will likely affect whether they endorse care as appropriate for humans to direct towards AI. As such, these kinds of individual differences may be important to consider when evaluating the cognitive mechanisms underpinning human–human and human–AI relational norms.

The research design described above can be adapted to further address mechanism: For example, researchers might consider manipulating the way that AI systems are themselves described (e.g., in terms of current or predicted future capacities, anthropomorphic qualities). Alternatively, researchers might depart from the self–report metrics described above, instead implementing various behavioral measures to test people's applications of relational norms to human–AI cases (e.g., gauging how people may speak differently with a human versus an AI companion). These kinds of comparisons across measures could clarify whether people's explicit beliefs about how relational norms should govern human–AI relationships align with their implicit attitudes and actual behavior.

Taken together, we see this kind of empirical project as an initial step towards understanding (and perhaps even forecasting) the nature of human–AI relationships. We see this as essential for informing the design and governance of AI systems that are more trustworthy, ethical, and aligned with human values.

4 Discussion and looking ahead

In this paper, we sketched how relational norms may differ between human–human and human–AI relationships across a range of contexts. We anticipate that some relational norms will transfer between human–human and human–AI relationships, whereas others will not. At present, we expect that people may be largely resistant to norms endorsing (costly) care–based behavior or mating–related behavior in human–AI relationships, although perhaps to different degrees depending on the relationship (e.g., stronger norms against flirtatious behavior when the AI is programmed to mimic a supervisory relationship than when it is programmed to serve as a 'romantic' companion). People may also resist being under the authority of an AI, complicating social roles characterized by hierarchy (e.g., between a manager and an employee). Indeed, people may even fear AI occupying these kinds of social roles [18] or reject that they should be in them altogether. We argue that understanding these nuances—by applying a relational norm framework that considers human–AI interaction in the context of specific social roles—is an important open direction for research in AI ethics and cognitive science, more broadly.

There is also a possibility that human–human relational norms will shift as social interaction with AI becomes increasingly common. For example, based on their increasing familiarity with AI systems in particular roles, will humans come to expect "one-way" relationships with other humans, where only one partner is acknowledged to have any needs? Will human–human relationships typically characterized by asymmetrical hierarchy norms, such as the teacher–student relationship, become more egalitarian? By studying human–human and human–AI relational norms (and, ideally, behavior) longitudinally, we can trace how these social landscapes may change over time.

Evaluating these norms longitudinally can also inform AI design. When people maintain similar relational expectations between human–exclusive and human–AI relationships, they may be more amenable to interacting with AI as a social partner, or regard those interactions as more authentic. We speculate that, at present, these expectations may be most similar in the case of common, relatively low–stakes AI social roles, such as AI assistants. But as AI models become more intelligent, capable, and humanlike in their presentation, people's expectations regarding which social roles are appropriate for AI may evolve accordingly. Social roles for AI that currently seem far–fetched (or even terrifying, to some), such as AI as long–term romantic partner or employer, may become more plausible—and more prevalent—as these systems improve.

In addition, we see this research program as potentially highlighting unintended consequences or risks of applying human–like relational norms to AI systems, such as the risk of undue anthropomorphism or emotional manipulation [24]. By observing trends in relational norms over time, we may better predict how people will engage with future AI models, enabling us to anticipate and hopefully avoid certain types of risk. This would require developers to draw on these insights to inform responsible AI design.

Our research program will also have significant implications for interpreting or adapting AI regulation (in both the present and in the future). For example, the EU AI Act takes a risk–based approach to AI governance, classifying entire categories of AI systems as "high– or low–risk" by area of application. Thus, all AI systems applied for medical or educational uses are automatically considered high risk.

Yet, it is not only the area of application that is relevant for determining risk—some uses of AI in medicine or education will be more risky than others. Future regulations informed by, and sensitive to, human–AI relational norms are likely to do a better job of identifying and addressing such risks.

Similarly, much soft law and scholarship in the area of AI ethics focuses on principles—such as safety, transparency, and privacy—which the design, development, and use of AI systems should respect [35, 37, 51]. However, as mentioned previously, how these abstract notions should be interpreted will depend on the socio–relational role the AI system is occupying and the way humans interact with it. Thus, a better understanding of human–AI relational norms may be one way of getting beyond abstract debates about principles and closer to addressing practical problems with direct impacts on human users of AI systems.

In addition to studying these norms longitudinally, we also recommend that this research be conducted cross–culturally. Advances in artificial intelligence may be not only transformative but radically transformative, leading "to societal change comparable to that precipitated by the agricultural or industrial revolutions" [30, p. 4]. These kinds of transformations affect the entire world, including future generations. Having a comprehensive understanding of human–AI relational norms for individuals of different backgrounds (e.g., within and outside of "WEIRD" cultures; [1, 2]) may be an essential step towards designing AI systems that operate properly within (i.e., with due sensitivity to) diverse socio–cultural contexts.

Importantly, relational norms or associated cooperative expectations may not only differ across countries/cultures but also within them (e.g., by demographic factors, such as age, previous experience with technology, previous experience with AI). These differences may have important implications for how AI systems are designed in light of relational norms. For example, AI systems designed to satisfy or adhere to the care function may have to exhibit very different types of caring behavior towards children, the elderly, and middle–aged adults.

In this paper, we have focused on how a better understanding of human–AI relational norms could be used to improve governance and use of AI systems. However, we recognize that this information, like conversational AI systems more generally, may be subject to dual–use risks [38]. Understanding how humans relate to AI systems differently based on social role could be harmfully misused. For example, AI systems involved in human–AI care–based interactions could conceivably be designed to elicit personal information from the human relational partner, which could then be misused in various ways [31]. One straightforward example would be eliciting financial information, such as passwords or credit card numbers; other examples include eliciting sensitive personal or medical information that could be used for targeted advertising, further model training, or sold to third parties [39]. Similarly, understanding how humans apply hierarchy–based norms to AI could be exploited to create AI systems that unduly influence human decision–making or behavior (e.g., in political contexts). Finally, there is also the risk that detailed knowledge of human–AI relational norms could be used to create more convincing deepfakes or AI impersonation systems. By mimicking the relational norms that humans expect in specific social contexts, malicious actors could create AI systems that are increasingly effective at deceiving humans in social engineering attacks or by spreading misinformation [25]. We see these as significant concerns, but we also note that understanding the dual–use potential of increased knowledge about human–AI relational norms is itself necessary to combat nefarious use cases.

While we are convinced that our proposed research program will provide valuable descriptive insights into human–AI relational norms, it is important to be clear about the distinction between descriptive findings and normative conclusions. Empirical data can inform, but cannot on their own determine, ethical prescriptions. The type of research we propose here offers a descriptive foundation upon which future normative work can build: We cannot make properly nuanced arguments about how things should be without first knowing how they are. AI systems have the potential to either reinforce or challenge existing norms, raising important ethical questions about which norms should be preserved or modified as human–AI interactions become more prevalent. However, again, such questions cannot be answered in the abstract. They require the type of empirical work we propose here.

Moreover, it is crucial to recognize that relational norms are not static but can change over time, often in response to evolving social values and practices. A prime example of this is the shift in the doctor–patient relationship over the past century.[6] This relationship has moved from a paternalistic model, where doctors made decisions with little patient input, to one emphasizing patient autonomy and shared decision–making. What was once seen primarily through a lens of hierarchical "care" has incorporated stronger elements of patient self–determination. Such changes over time illustrate that while empirical data on cooperative functions in human–AI relationships are necessary, they do not themselves fully determine the extent to which these functions are normatively desirable.

To navigate the complex interplay between empirical findings and normative conclusions in this domain, we and other researchers can employ methodologies that bridge the

[6] We thank an anonymous reviewer for this key point and example, which we adopt with gratitude.

descriptive–normative gap. Earp et al. [20] propose several strategies for leveraging empirical data in the service of normative arguments within experimental philosophical bioethics. These strategies include assigning *prima facie* normative weight to widely–shared moral judgments among a group of stakeholders, if these judgments can be shown to be robust and appropriately responsive to normative reasons; using "debunking" strategies to subtract normative weight from moral judgments that are fickle or inappropriately responsive to normatively (ir)relevant factors (e.g., framing effects, but see [16]); and assigning comparable normative weight to diverse views among different groups of stakeholders, if these views can be shown to track different, yet reasonable underlying ethical commitments (i.e., adopting a policy of pluralism). All of these could be applied to research into human–AI relational norms. Complementing this approach, Savulescu et al. [50] propose and defend the Collective Reflective Equilibrium in Practice (CREP) procedure, which integrates empirical data on stakeholder preferences and expert judgments with ethical principles to generate justified policy recommendations.

Ultimately, the empirical research program proposed here must be complemented by both "armchair" ethical theorizing in philosophy, as well as the types of integrative methodologies just mentioned. Nevertheless, we argue that it is crucial for all of these approaches to inform each other, and to be based on a rich descriptive understanding of the relational norms that people in different socio–cultural contexts do, in fact, hold—even if it is ultimately argued that some of these norms have undesirable consequences (or are intrinsically problematic) and should therefore be reformed or abandoned.

5 Conclusion

In the present work, we have argued that there may be considerable value in understanding how people represent (and distinguish between) human–human and human–AI relational norms. As AI systems advance and continue to assume socio–relational roles traditionally filled by humans, it will be critical to understand how these developments may complicate people's cooperative expectations about relationships—whether that be with other humans, or, indeed, with AI. Given the rapid pace of AI development, we think it is crucial to explore human–AI relational norms and translate these findings into contextually–sensitive AI governance and design. We expect that changes in human–AI relational norms will track with advances in AI models; we therefore urge policy–makers and scholars in this area to help facilitate and keep abreast of developments in research on, and understanding of, human–AI relational norms.

Relationship–specific cooperative norms governing human–AI interactions represent a critical and largely unexplored frontier in AI ethics, governance, and design. By examining how people apply, adapt, or resist applying human–human relational norms in human–AI contexts, we can inform more nuanced AI development, governance, use, and integration strategies. The potential implications of this work are far–reaching, from improving the design of AI systems to informing more contextually–sensitive and effective AI regulations. Moreover, by tracking how these norms evolve over time and across cultures, we can better understand, prepare for, and influence the long–term societal impacts of widespread human–AI interaction. We call upon researchers across disciplines, policymakers, and AI developers to engage with this crucial area of study. Only through collaborative, interdisciplinary efforts can we hope to navigate the complex landscape of human–AI relationships.

Acknowledgements Any use of generative AI in this manuscript adheres to ethical guidelines for use and acknowledgement of generative AI in academic research [53]. Specifically, Anthropic's Claude 3.5 Sonnet was used during the writing of this manuscript to edit and shorten some passages of the authors' own written text. Each author has made a substantial contribution to the work, which has been thoroughly vetted for accuracy, and assumes responsibility for the integrity of their contributions.

Funding This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: -AISG3-GV-2023-012), by the National Institute for Health and Care Research (NIHR) Oxford Health Biomedical Research Centre (Award NIHR203316), by the Wellcome Trust (Wellcome Centre for Ethics and Humanities 203132/Z/16/Z), and by the Novo Nordisk Foundation grant for a scientifically independent International Collaborative Bioscience Innovation & Law Programme (Inter-CeBIL Programme, Grant no. NNF23SA0087056). The views expressed are those of the authors and do not necessarily reflect those of the funding bodies or by the Department of Health and Social Care.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Atari, M., Haidt, J., Graham, J., Koleva, S., Stevens, S.T., Dehghani, M.: Morality beyond the WEIRD: how the nomological network of morality varies across cultures. J. Pers. Soc. Psychol. **125**(5), 1157–1188 (2023). https://doi.org/10.1037/pspp0000470

2. Atari, M., Xue, M.J., Park, P.S., Blasi, D., Henrich, J.: Which humans? (2023). https://osf.io/5b26t
3. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623 (2021). https://doi.org/10.1145/3442188.3445922
4. Biden, J.R.: Executive order on the safe, secure, and trustworthy development and use of artificial intelligence (2023). https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/
5. Bigman, Y.E., Wilson, D., Arnestad, M.N., Waytz, A., Gray, K.: Algorithmic discrimination causes less moral outrage than human discrimination. J. Exp. Psychol. Gen. **152**(1), 4–27 (2023). https://doi.org/10.1037/xge0001250
6. Bloom, P.: Family, community, trolley problems, and the crisis in moral psychology. Yale Rev. **99**(2), 26–43 (2011). https://doi.org/10.1111/j.1467-9736.2011.00701.x
7. Bostrom, N.: Superintelligence: Paths, Dangers, Strategies. Oxford University Press, Oxford (2014)
8. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M.T., Zhang, Y.: Sparks of artificial general intelligence: early experiments with GPT-4 (2023). https://doi.org/10.48550/arXiv.2303.12712
9. Bugental, D.B.: Acquisition of the algorithms of social life: a domain-based approach. Psychol. Bull. **126**(2), 187–219 (2000). https://doi.org/10.1037/0033-2909.126.2.187
10. Chalmers, D.J.: Could a large language model be conscious? (2023). https://doi.org/10.48550/arXiv.2303.07103
11. Clark, M.S., Mills, J.: The difference between communal and exchange relationships: what it is and is not. Pers. Soc. Psychol. Bull. **19**(6), 684–691 (1993). https://doi.org/10.1177/0146167293196003
12. Coeckelbergh, M.: Technology games/gender games. from Wittgenstein's toolbox and language games to gendered robots and biased artificial intelligence. In: Feminist Philosophy of Technology, pp. 27–38. Springer (2020). https://doi.org/10.1007/978-3-476-04967-4_2
13. Colombatto, C., Fleming, S.M.: Folk psychological attributions of consciousness to large language models. Neurosci. Conscious. **2024**(1), niae013 (2024). https://doi.org/10.1093/nc/niae013
14. Curry, O.S., Jones Chesters, M., Van Lissa, C.J.: Mapping morality with a compass: testing the theory of 'morality-as-cooperation' with a new questionnaire. J. Res. Pers. **78**, 106–124 (2019). https://doi.org/10.1016/j.jrp.2018.10.008
15. Danaher, J.: Toward an ethics of AI assistants: an initial framework. Philos. Technol. **31**(4), 629–653 (2018). https://doi.org/10.1007/s13347-018-0317-3
16. Demaree-Cotton, J.: Do framing effects make moral intuitions unreliable? Philos. Psychol. **29**(1), 1–22 (2016). https://doi.org/10.1080/09515089.2014.989967
17. Depounti, I., Saukko, P., Natale, S.: Ideal technologies, ideal women: AI and gender imaginaries in Redditors' discussions on the Replika bot girlfriend. Media Cult. Soc. **45**(4), 720–736 (2023). https://doi.org/10.1177/01634437221119021
18. Dong, M., Conway, J., Bonnefon, J.-F., Shariff, A., Rahwan, I.: A psychological model predicts fears about artificial intelligence across 20 countries and 6 domains of application (2023). https://osf.io/preprints/psyarxiv/pjvqt
19. Earp, B.D.: Relational morality. Doctoral dissertation, Yale University (2021). https://elischolar.library.yale.edu/gsas_dissertations/325
20. Earp, B.D., Lewis, J., Dranseika, V., Hannikainen, I.R.: Experimental philosophical bioethics and normative inference. Theor. Med. Bioeth. **42**(3), 91–111 (2021). https://doi.org/10.1007/s11017-021-09546-z
21. Earp, B.D., McLoughlin, K.L., Monrad, J.T., Clark, M.S., Crockett, M.J.: How social relationships shape moral wrongness judgments. Nat. Commun. **12**(1), 5776 (2021). https://doi.org/10.1038/s41467-021-26067-4
22. European Parliament. EU AI Act: First Regulation on Artificial Intelligence (2024). https://artificialintelligenceact.eu/the-act/
23. Firt, E.: The missing G. AI Soc. **35**(4), 995–1007 (2020). https://doi.org/10.1007/s00146-020-00942-y
24. Gabriel, I., Manzini, A., Keeling, G., Hendricks, L.A., Rieser, V., Iqbal, H., Tomašev, N., Ktena, I., Kenton, Z., Rodriguez, M., El-Sayed, S., Brown, S., Akbulut, C., Trask, A., Hughes, E., Bergman, A.S., Shelby, R., Marchal, N., Griffin, C., Isaac, W., Manyika, J.: The ethics of advanced AI assistants (2024). https://doi.org/10.48550/arXiv.2404.16244
25. Gallagher, S., Gelman, B., Taoufiq, S., Vörös, T., Lee, Y., Kyadige, A., Bergeron, S.: Phishing and social engineering in the age of LLMs. In: Kucharavy, A., Plancherel, O., Mulder, V., Mermoud, A., Lenders, V. (eds.) Large Language Models in Cybersecurity: Threats, Exposure and Mitigation, pp. 81–86. Springer Nature, Switzerland (2024). https://doi.org/10.1007/978-3-031-54827-7_8
26. Gao, S., He, L., Chen, Y., Li, D., Lai, K.: Public perception of artificial intelligence in medical care: content analysis of social media. J. Med. Internet Res. **22**(7), e16649 (2020). https://doi.org/10.2196/16649
27. Goertzel, B.: The Hidden Pattern: A Patternist Philosophy of Mind. BrownWalker Press, Irvine (2006)
28. Good, I.J.: Speculations Concerning the First Ultraintelligent Machine. Elsevier, Amsterdam (1966)
29. Gray, H.M., Gray, K., Wegner, D.M.: Dimensions of mind perception. Science **315**(5812), 619 (2007). https://doi.org/10.1126/science.1134475
30. Gruetzemacher, R., Whittlestone, J.: Defining and unpacking transformative AI (2019). arXiv:1912.00747v1
31. Gumusel, E., Zhou, K. Z., Sanfilippo, M.R.: User privacy harms and risks in conversational AI: a proposed framework (2024). arXiv:2402.09716
32. Haidt, J., Baron, J.: Social roles and the moral judgement of acts and omissions. Eur. J. Soc. Psychol. **26**(2), 201–218 (1996)
33. Hamilton, V.L., Sanders, J.: The effect of roles and deeds on responsibility judgments: the normative structure of wrongdoing. Soc. Psychol. Q. **44**(3), 237–254 (1981). https://doi.org/10.2307/3033836
34. Hester, N., Gray, K.: The moral psychology of raceless, genderless strangers. Perspect. Psychol. Sci. **15**(2), 216–230 (2020). https://doi.org/10.1177/1745691619885840
35. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**(9), 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2
36. Kempt, H.: Synthetic friends. In: Synthetic Friends, pp. 111–161. Springer International Publishing, Berlin (2022)
37. Khan, A.A., Badshah, S., Liang, P., Khan, B., Waseem, M., Niazi, M., Akbar, M.A.: Ethics of AI: a systematic literature review of principles and challenges (2021). https://doi.org/10.48550/arXiv.2109.07906
38. Koplin, J.J.: Dual-use implications of AI text generation. Ethics Inf. Technol. **25**(2), 32 (2023). https://doi.org/10.1007/s10676-023-09703-z
39. Kumar, V., Ashraf, A.R., Nadeem, W.: AI-powered marketing: what, where, and how? Int. J. Inf. Manag. **77**, 102783 (2024). https://doi.org/10.1016/j.ijinfomgt.2024.102783
40. Lee, M.K.: Understanding perception of algorithmic decisions: fairness, trust, and emotion in response to algorithmic

management. Big Data Soc. **5**(1), 2053951718756684 (2018). https://doi.org/10.1177/2053951718756684
41. Legg, S., Hutter, M.: A formal measure of machine intelligence (2006). arXiv:cs/0605024
42. Makovi, K., Sargsyan, A., Li, W., Bonnefon, J.-F., Rahwan, T.: Trust within human-machine collectives depends on the perceived consensus about cooperative norms. Nat. Commun. **14**(1), 3108 (2023). https://doi.org/10.1038/s41467-023-38592-5
43. Mallick, R., Flathmann, C., Lancaster, C., Hauptman, A., McNeese, N., Freeman, G.: The pursuit of happiness: the power and influence of AI teammate emotion in human-AI teamwork (2023). https://www.tandfonline.com/doi/abs/10.1080/0144929X.2023.2277909
44. Maria, K., Drigas, A., Skianis, C.: Chatbots as cognitive, educational, advisory & coaching systems education. Tech. Soc. Sci. J. **30**, 109–126 (2022)
45. Mills, J., Clark, M.S.: Communal and exchange relationships: controversies and research. In: Theoretical Frameworks for Personal Relationships. Psychology Press, Hove (1994)
46. Müller, V.C., Bostrom, N.: Future progress in artificial intelligence: a survey of expert opinion. In: Müller, V.C. (ed.) Fundamental Issues of Artificial Intelligence, pp. 555–572. Springer International Publishing, Berlin (2016). https://doi.org/10.1007/978-3-319-26485-1_33
47. Pauketat, J.V.T., Anthis, J.R.: Predicting the moral consideration of artificial intelligences. Comput. Hum. Behav. **136**, 107372 (2022). https://doi.org/10.1016/j.chb.2022.107372
48. Porsdam Mann, S., Earp, B.D., Nyholm, S., Danaher, J., Møller, N., Bowman-Smart, H., Hatherley, J., Koplin, J., Plozza, M., Rodger, D., Treit, P.V., Renard, G., McMillan, J., Savulescu, J.: Generative AI entails a credit-blame asymmetry. Nat. Mach. Intell. **5**(5), 472–475 (2023). https://doi.org/10.1038/s42256-023-00653-1
49. Rai, T.S., Fiske, A.P.: Moral psychology is relationship regulation: moral motives for unity, hierarchy, equality, and proportionality. Psychol. Rev. **118**(1), 57–75 (2011). https://doi.org/10.1037/a0021867
50. Savulescu, J., Gyngell, C., Kahane, G.: Collective reflective equilibrium in practice (CREP) and controversial novel technologies. Bioethics **35**(7), 652–663 (2021). https://doi.org/10.1111/bioe.12869
51. UNESCO. Recommendation on the ethics of artificial intelligence (2021). https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence
52. Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L.A., Rimell, L., Isaac, W., Irving, G., Gabriel, I.: Taxonomy of risks posed by language models. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 214–229 (2022). https://doi.org/10.1145/3531146.3533088
53. Porsdam Mann, S., Vazirani, A.A., Aboy, M. et al.: Guidelines for ethical use and acknowledgement of large language models in academic writing. Nat. Mach. Intell. 6, 1272–1274 (2024). https://doi.org/10.1038/s42256-024-00922-7

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

AI and Ethics
https://doi.org/10.1007/s43681-024-00628-x

ORIGINAL RESEARCH

Waging warfare against states: the deployment of artificial intelligence in cyber espionage

Wan Rosalili Wan Rosli[1]

Received: 1 July 2024 / Accepted: 16 November 2024
© Crown 2025

Abstract
Cyber espionage has significantly been viewed as a risk towards nation-states, especially in the area of security and protection of Critical National Infrastructures. The race against digitisation has also raised concerns about how emerging technologies are defining how cyber activities are linked to waging warfare between States. Real-world crimes have since found a place in cyberspace, and with high connectivity, has exposed various actors to various risks and vulnerabilities, including cyber espionage. Cyber espionage has always been a national security issue as it does not only target States but also affects public–private networks, corporations and individuals. The challenge of crimes committed within the cyber realm is how the nature of cybercrimes distorts the dichotomy of state responsibility in responding to cyber threats and vulnerabilities. Furthermore, the veil of anonymity and emerging technologies such as artificial intelligence have further provided opportunities for a larger scale impact on the state for such crime. The imminent threat of cyber espionage is impacting the economic and political interactions between nation-states and changing the nature of modern conflict. Due to these implications, this paper will discuss the current legal landscape governing cyber espionage and the impact of the use of artificial intelligence in the commission of such crimes.

Keywords Cyber espionage · Artificial intelligence · Cyber operations · Warfare · International law

1 Introduction

Cyber espionage is coined as a type of cyber warfare that includes disruptions, surveillance, damage, and deletion of information. Cyber espionage threats are powerful enough to immobilise a state and disrupt the running of critical national infrastructures, where the sabotage of one sector may result in total system failure, data leakage, and even system harm. If one sector fails, it would cause a domino effect on other critical sectors which could result in a national catastrophe. The European Union Agency for Cybersecurity (ENISA) defines it as a type of threat committed by using a computer network to gain illicit access to confidential information usually held by state governments or organisations [1]. Such information or intelligence refers to information which is of military, political and economic value. The United States Central Intelligence Agency (CIA) contended that intelligence involves knowledge and foreknowledge of the world which drives decisions by policymakers. The Tallinn Manual, which is one of the first documents that outlines the rules for nation-state cyber warfare has stated that espionage as an act undertaken clandestinely or under false pretences that uses cyber capabilities to gather or attempt to gather information [2].

Previously, espionage was usually done by states sending their most skilled agent into the territory of their physical adversaries. The development of technology has changed the way espionage is conducted where states now spy on their enemies by using more sophisticated methods such as drones and spy balloons. For years, state and non-state actors have been engaging in spying to gain information, assess potential threats, and prepare for attacks [3]. Essentially, throughout history, states have continued to engage in clandestine activities against each other through spying and sabotaging [4]. This type of practice has long been viewed as an attribute of warfare, state intelligence and combatant engagement. The race towards digitalisation and the emergence of disruptive technologies are now viewed to be an

✉ Wan Rosalili Wan Rosli
rosalili2301@gmail.com; w.r.wanrosli@bradford.ac.uk

[1] School of Law, Faculty of Management, Law and Social Sciences, University of Bradford, Bradford, UK

Published online: 08 January 2025

expansion of the term cyber warfare. This also meant that nation-states began to move their focus to the protection of sectors within the critical national infrastructure which required special cyber protection and security.

In combination with the growing sophistication of cyber actors and activities, it further expands opportunities for coordinated and advanced attacks that can speedily disrupt the operational efficiency of designated critical national infrastructure e.g., the operation of the financial markets, electricity, security infrastructure and general elections [5]. When cyber espionage is part of a broader military campaign it can also lead to disruption of public infrastructure and property [6]. Cyber espionage attacks can be motivated by monetary value or may also be deployed in conjunction with military operations, or as an act of cyber terrorism or cyber warfare.

Espionage has become an imminent threat and risk to nations, especially in understanding how technology shares and influences the relationship between states. Many nations and international bodies around the world have created their own definition of cyber espionage and it has been proven to be difficult to reach a consensus due to the secretive nature of espionage and its relation to warfare. On the other hand, states have also argued that cyber espionage tactics may be used by intelligence services to protect national security by preventing terrorist attacks [7]. It can also be used in war where the information collected from the enemy can be used successfully to gain warfare advantage [8]. With the realization of the opportunities that cyberspace offers, governments across the world are building on cyber-forces, which are tasked with accomplishing their national cybersecurity policy goals and intelligence agendas [9].

Against this backdrop, the implications for international security are serious and far-reaching. As nation-states contend for dominance in cyberspace and geopolitical tensions escalate, the risk of Artificial Intelligence (AI) -enabled cyber activities will unarguably exacerbate tensions. The potential for misattributed attacks, unintended escalation, and the increase of cyber operations threatens to destabilise the balance of power and heighten the risks of conflict escalation [10]. It is imperative to therefore identify the emergent contours of warfare in the digital age and develop sound strategies for safeguarding international security.

2 International law and the waging of wars between states

Espionage has always been a feature of interstate relations deeply rooted in secret intelligence-gathering activities conducted between states. Article 2 of the 1945 Charter of the United Nations (UN Charter) explicitly explained that all member states should refrain from using force or threats that encroach on another state's territorial integrity or political independence. Despite the availability of an international framework for the governance of espionage, specific treaties and conventions addressing the intrusion are notably absent. The commission of espionage however is dealt with under principles of international humanitarian law, sovereignty and customary practices.

The concept of espionage also violates the duty of non-intervention which was also mandated under the UN Charter. Jamnejad and Wood [11] explains that the principle of non-intervention includes the right to sovereignty, territorial integrity and political independence of every State. The principle has also been accepted to apply to cyber operations. However, the non-intervention principle only applies to States as it is an inter-state doctrine. Any activities conducted by non-state actors do not apply unless through the doctrine of attribution in international law. Within the context of state-sponsored cyber operations, which target individuals and companies, the non-intervention principle does not apply as it does not affect the state's exclusive rights to sovereign functions [12]. To date, in so far cyber operations such as cyber espionage, states have only responded to attacks violating the states' ability to exercise their sovereign powers [12]. This can be seen by the response of the Not-Petya attack targeting Ukraine's critical national infrastructures which was labelled as an act disregarding Ukrainian sovereignty as appose to the spear fishing attack campaign aimed at private universities, companies and NGOs by Iran which was not to be said a violation of international law [13].

The principle of non-intervention and its requirement of fulfilling the element of violation of sovereignty proves challenging as the nature of cyberspace which allows covert, anonymous and remote operations which can also go beyond borders. Furthermore, careful investigations and cooperation between international partners are crucial in order to track and find the perpetrators. Recent attacks have shown more coercive and intrusive patterns where these intrusions often target the critical national infrastructure of States such as transportation, energy and manufacturing [13]. However, Oorsprong [14] highlighted that espionage could hardly be considered as an international crime as underlying reasons for collecting the information would usually not constitute to interference or violate the non-intervention principle.

Regardless of the argument of how cyber operations such as cyber espionage would go against international law principles, the majority of states have always viewed espionage-like acts to be an important element in maintaining peace, stability and security between nation-states. Interestingly, Hernandez [15] and Scott [16] have also highlighted that the practice of espionage between states is legal on the basis

of anticipatory or pre-emptive self-defence. Such belief is derived from the power of knowing the ability and strength of other states can be determined, and the balance of power and status quo can be achieved which results in non-confrontational behaviour [17].

Apart from nation-states engaging in espionage, non-state actors also have the ability to take part in espionage activities. However, nation-states will always play a central role in conflicts among states due to the financial, technological and cyber capabilities to conduct sophisticated operations and share geopolitical dynamics in cyberspace. Rauta [17] contends that these proxy groups and non-state actors engage in warfare through sponsorship of the state to prevent attribution. These entities would carry out covert activities, data mining, and cyber-attacks to advance the strategic interest on behalf of the sponsoring state. It must be noted that actions or behaviour which could be triggering an act of war would be instances where unfriendly economic or commercial actions or any behaviour that insults national price. However, acts of cyber espionage are stealthier rather than traditional warfare which is more transparent. This in turn blurs the line of retaliation and response due to the targets are usually private companies and not directly towards the State.

In another sphere, the dichotomy between developed and developing nations adds another division to the conversation. Spectres of superiority and dominance between states have contributed to the ongoing geopolitical and cyber-diplomatic struggles globally. Superiority is being determined by one nation's economic, political and military strength which has led to conversations and debates on cyber governance [18]. States such as China and Russia have resorted to diverse covert activities to gain economic and technological power. The advancement of technology has broken traditional barriers, and the cyberspace environment now offers states the opportunity to take advantage of such cyber capabilities [19]. This has inherently sparked the need for a unified framework to govern cybercrime in the form of the world's first international cybercrime treaty [20].

3 Proliferation of artificial intelligence (AI) in cyber espionage

The convergence of law and technologies has brought about various implications and challenges. The use of emerging technologies afforded by AI has proven to be the holy grail in combatting crimes within the real world and cyberspace. It also has become an integral part of society, driving innovations across sectors including legal or illegal cyber operations. The deployment of AI in the fight against cybercrime has also facilitated the investigation of crimes committed in cyberspace, especially for entities that hide behind the shadows of cyberspace and their digital footprints. The Swiss Federal Intelligence Service (FIS) Annual Situation Report (2023) highlighted that espionage is driven by various motives which involve various aims such as economic and political superiority. The FIS also contended that the availability of new emerging technologies acts as a sandbox for various actors to develop new solutions to problems including the need for states to focus on lingering military conflict [21].

The use of AI aids in detecting cyber espionage including the use of biometric recognition technologies such as facial detection to sift through data available online and provide immediate notification [22]. The emergence of Generative AI (GenAI) has completely changed the landscape of commission and protection against cyber espionage. GenAI has revolutionised tools and solutions used by States to safeguard their systems from infiltration [23]. Such technology provides defence against automation, cybersecurity reporting, threat intelligence, identification of attacks, incidence response and malware detection. Apart from that, GenAI has also unleashed both known and unknown cyberattack vectors in the form of automated hacking, code-adapting malware and AI-crafted phishing e-mails.

The increased dependency on AI has provided law enforcement agencies with the tools to fight cybercrime more effectively. New emerging technologies such as build-in cyber espionage search engines which eliminates cumbersome elements of extracting information-laden threats and discovering digital footprints of offenders in order to link the relevant information which is time-consuming [23]. The use of such technology has broken barriers of time-consuming work with the capability of shifting through unlimited volumes of data, the least amount of time and minimal resources to track down leads for any vulnerabilities and threats against the State [23]. This has also led to the eradication of threats of violence and improve security measures within the State.

The duality impact of AI must also be recognised. AI has proven to be lethal, especially in coordinating cyber-attacks via sophisticated algorithms that can penetrate even the most complex security systems, data theft of sensitive information such as state secrets or military capability information, and disruption of multisector services [24]. Samuel-Okon et al. [25] contend that AI-driven malware can evade traditional detection methods and cause extensive damage as it has the capability to adapt and evolve causing widespread data breaches. AI algorithms can analyse a vast amount of data derived from various sources such as public records, network traffic and social media in order to identify potential targets. These algorithms can also develop stealthy tools to infiltrate and bypass security defences and

intrusion detection systems. Moisset [26] contends that AI algorithms can assist in prioritising targets based on factors such as potential intelligence value, susceptibility to exploitation and significance. Catalano et al. [27] highlighted that adversarial machine learning also enables attackers to evade signature-based detection.

AI is no longer confined to science fiction but is now a critical component of military strategy. Nations are actively integrating AI into their defense capabilities, leading to a paradigm shift in how wars are fought. In recent years, characterized by rapid technological advancement, the emergence of AI and cyber espionage has become a defining feature of modern warfare and international security [28]. As nations increasingly harness the power of AI to bolster their defense capabilities and conduct intelligence operations, the landscape of conflict has evolved to another dimension—one where states use various means to try and gather information from other states without consent [28].

The emergence of GenAI has also changed the landscape of the commission and protection against cyber espionage. GenAI-aided attacks have now unleashed both known and unknown cyberattack vectors. Gupta *et. al* highlighted that GenAI acts as a double-edged sword in cybersecurity as it facilitates and benefits both sides of the coin [28]. Nation states have now moved to GenAI tools to safeguard their systems from infiltration. Information gathered from Large Language Models (LLM) trains intelligence data such as risk, vulnerabilities and prediction of attacks [28]. States now have the capability to process large information to enhance their intelligence response and even predict emerging and future threats [29]. Furthermore, such advanced technology is also important in developing effective ethical guidelines for strengthening cyber defense within a system [30]. According to Gupta, the use of GenAI provides support in cyber defense automation, cybersecurity reporting, threat intelligence, identification of cyber-attacks, incident response guidance and even malware detection [30].

The worldwide movement to govern the use of AI has also impacted the way the State responds to the use of GenAI in the commission of cyber espionage. The legal guidelines and practices such as the EU AI Act aim to give clear guidelines on the misuse and responsibility of users when using such technology. Such moves have leveraged the use of such technology and have also pushed states to give clear guidelines on the use of such emerging technology.

4 The legal response to cyber espionage

Effective governance of cybercrimes and cyber operations have been at the forefront of every nation, especially in eradicating national security threats. The Office of the National Counterintelligence Executive has also highlighted that cyber espionage heightens the risk to international peace and security [31]. The United Nations Special Rapporteur on Counterterrorism and Human Rights have also called for a legal response to spyware proliferation. However, it must be noted that the line between lawful and unlawful cyber espionage or intrusions remains unclear. With the current economic and political global landscape, states have now turned to engage in diverse cyber activities on the pretext of national security [32]. Apart from relying on traditional laws such as national legislation and statutes to govern cybercrimes, states have now moved to various other mechanisms in addressing cybersecurity matters such as devising cybersecurity strategies in the form of state policies in order to maintain cyber resilience across sectors [32]. A cybersecurity strategy uses a top-down approach which contains objectives and protocols in order to protect a state's critical national infrastructure. The strategy is designed to maximise the security and resilience of a nation by outlining the duties and obligations of major stakeholders within the state ecosystem [33]. States also view that having a strategy in place is a proactive approach to mitigate the risks and threats of cyber-attacks. Furthermore, such strategies also recognises the ever-evolving advancement of technologies and continuously devise ways for the state to adapt and align with national security.

The Assange and Snowden leak has revealed the gravity of intrusion and data breach to states which motivates them to seek a harmonised approach to regulate and mitigate the risk of cyber espionage. States have now moved towards the curtailing of remote access cyber espionage by deploying state-of-the-art technologies to restrict such operations. In 2023, aware of the risk on infiltration and the risk of espionage to national security, the United States passed the Presidential Initiative for Democratic Renewal which intends to control the misuse of technology and digital authoritarianism [32]. The Initiative also aims at combatting digital domination which includes an executive order to prohibit the use of spyware which poses a risk to national security. Apart from that, the United States also enacted the Espionage Act which restricts the access and disclosure of information which can be used against the United States or provide undue advantage to a foreign state which would compromise national security [34]. The former United States National Security Agency contractor, Edward Snowden was charged under this Act, but has yet to face trial due to being granted the right of asylum in Russia.

Traditional legislation governing cyber espionage can be found in every country in one form or another. In China, cyber espionage is governed under the National Intelligence Law 2017 which aims to strengthen and safeguard national intelligence work and preserve state security and interest

[35]. Interestingly, the law also attempts to balance the securitization of sectors to the protection of human rights and the interest of individual citizens. In Russia, the Criminal Code of the Federation, Law NO. FZ-190 criminalise espionage and it is seen as treason. The law includes acts of assisting foreign states, disclosure of national secrets or anything that would be detrimental to the security of the Russian Federation [36].

The European Union has also taken steps to regulate cyber espionage when recently the European Parliament called for a stricter approach to the use and abuse of spyware. Furthermore, in February 2024, the United Kingdom and France, hosted nation states, regional organizations and other stakeholders in an inaugural conference to tackle the proliferation and irresponsible use of commercial cyber intrusion tools and services [37]. The conference also led to the signing of the cyber declaration—The Pall Mall Process: Tackling the Proliferation and Irresponsible Use of Commercial Cyber Intrusion Capabilities [38].

The Dutch Military Intelligence and Security Service (MIVD) highlighted that Western governments, defence companies and international organisations are now being targeted with cyber espionage [39]. MIVD also contended within it 2024 annual report that Chinese spies has recently targeted Dutch semiconductor, aerospace and maritime industries in order to gather intelligence in strengthening its armed forces. However, China have always routinely denied any form of allegations relating it to cyber espionage [40].

The first quarter of the year has also seen Western governments especially within Europe being under siege by cyberattacks and operations of espionage. The United Kingdom Ministry of Defence has also suffered cyberattacks which exposed the sensitive data of thousands of its troops [40]. The breach targeted the weakness of a third-party contractor who runs the payroll systems for the Ministry which leaked more than 270,000 data of former and current armed forces members. Germany was also targeted when a 38-min audio recording of four top military personnel discussing the support for Ukraine including using missiles to attack Russia [40].

The legal response to cyber espionage takes many forms and challenges. Apart from what has been discussed above, such attacks depend on the type of cyber operations and whether the implications are confined to the national level or have impacted internationally. For example, an Iranian-based group named Helix Kitten has been notorious for targeting companies or organisations within the sectors of transportation, finance, hospitality and telecommunications and using sophisticated methods via structured spear-phishing attacks to targeted personnel. Another national state actor, which is an organisation that is supported by a specific nation is a group labelled as Fancy Bear, a Russian-based group that is linked to the Russian Intelligence Service. Their primary aim is to target U.S. political organisations and European military organisations through zero-day vulnerabilities and malware drop websites disguised as new sources.

The response to such invasive actions can be both military and non-military. Non-military responses would include diplomatic measures such as demarches and expelling diplomats, legal actions such as cybercrime legislation and international legal cooperation, cyber defence and counterintelligence and also economic sanctions. However, the unequal level of digital capacity between nation-states also serves as a barrier to efforts of harmonisation of the law and better governance of cyber espionage. The challenges differ across regional boundaries considering limited resources, technical capabilities, and other social realities. A general lack of awareness of cybersecurity best practices also leaves many states vulnerable.

5 International security and cyber espionage governance

The ever-evolving threat of cyber espionage cannot be denied especially within the realm of cybersecurity with implications that could range from national security to political and economic stability. This type of cyber operation has been a common practice by states and non-state actors who practice recognisance activities to measure the strength and capability of other adversaries. From using old tactics such as sending agents into other jurisdictions to gather intelligence to the emergence of new technologies which transcend the recognition of borders and time, cyber espionage has time and time again proven to be a risk that needs to be managed.

Calls for criminalisation and clear classification of this cyber operation have led to efforts to harmonise international perspectives due to domestic laws being insufficient to address the challenges to national security, privacy and data protection and extraterritorial jurisdiction. The apparent global digital divide and access to justice for all must also be addressed. The best way forward would be to have a treaty to govern and clearly outline the use of emerging technology in the commission of cyber espionage. However, efforts to materialise an international treaty to govern such cybercrime pose a multitude of challenges due to the different interests and priorities of states especially in the area of sovereignty and security. Scholars have highlighted that a more realistic approach would be to completely prohibit cyber espionage and have in place norms and practices agreed upon by all states. Greiman (2018) further states that if States want non-binding responsibilities in cyberspace,

therefore their actions and practice must demonstrate their willingness to achieve that goal.

The increasing reliance on artificial AI by state intelligence would depend on the capability and operationalisation of intelligence activities. States will invest and focus on advancing their technological resources which will enhance their cyber power and international presence. This means that in the coming years, there will be an urgent need to nation states, international organizations and other stakeholders to consider a normative framework for the use of such capabilities. Furthermore, AI governance is vital as a way forward to invest in the development of AI resources. The passing of the AI Act has provided an opportunity for other states to also initiate their own strategies, policies and legal framework in governing AI and other emerging technologies. Apart from finding ways to regulate technologies, enhancing cybersecurity measures and deterring emergent threats are also critical in ensuring effective governance of cyber espionage. Collaborative intelligence is also essential to identify emerging risks and threats more effectively. Finally, by fostering cybersecurity awareness, investment in state-of-the-art defensive technologies and strengthening international cooperation, states can mitigate the risks of cyber espionage and create a more secure and trusted digital ecosystem.

Funding The author has no relevant financial or non-financial interests to disclose. No funding was received to assist with the preparation of this manuscript.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. ENISA Threat Landscape 2020—Botnet (2020). [Report/Study]. ENISA. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2020-botnet
2. Schmitt, M.N. (ed.): Tallinn manual on the international law applicable to cyber warfare. Cambridge University Press, Cambridge (2013)
3. Føllesdal, A.: The significance of state consent for the legitimate authority of customary international law. In: Merkouris, P., Kammerhofer, J., Arajärvi, N. (eds.) The Theory, Practice, and Interpretation of Customary International Law. The Rules of Interpretation of Customary International Law, pp. 105–136. Cambridge University Press, Cambridge (2022)
4. Fjäder, C.: The nation-state, national security and resilience in the age of globalisation. Resilience **2**(2), 114–129 (2014). https://doi.org/10.1080/21693293.2014.914771
5. Baker, M.S., Baker, J., Burkle, F.M., Jr.: Russia's hybrid warfare in Ukraine threatens both healthcare & health protections provided by international law. Ann. Glob. Health **89**(1), 1–6 (2023)
6. Harknett, R.J., Smeets, M.: Cyber campaigns and strategic outcomes. J. Strateg. Stud. **45**(4), 534–567 (2022). https://doi.org/10.1080/01402390.2020.1732354
7. Byman, D.: How to think about state sponsorship of terrorism. In: Survival: August 2023, pp. 101–121. Routledge (2023).
8. Byman, D., Wittes, B.: Reforming the NSA: how to spy after Snowden. Foreign Aff. **93**, 127 (2014)
9. Liebetrau, T.: Organizing cyber capability across military and intelligence entities: collaboration, separation, or centralization. Policy Des. Pract. **6**(2), 131–145 (2023). https://doi.org/10.1080/25741292.2022.2127551
10. Egloff, F.J., Smeets, M.: Publicly attributing cyber attacks: a framework. J. Strateg. Stud. **46**(3), 502–533 (2023). https://doi.org/10.1080/01402390.2021.1895117
11. Jamnejad, M., Wood, M.: The principle of non-intervention. Leiden J. Int. Law **22**(2), 345–381 (2009)
12. Moulin, T.: Reviving the principle of non-intervention in cyberspace: the path forward. J. Conflict Secur. Law **25**(3), 423–447 (2020)
13. Hüsch, P.: Non-intervention thresholds in cyberspace—in the shadow of the sovereignty debate? Nord. J. Int. Law **92**(3), 371–393 (2023)
14. Oorsprong, F., Ducheine, P., Pijpers, P.: Cyber-attacks and the right of self-defense: a case study of the Netherlands. Policy Des. Pract. **6**(2), 217–239 (2023). https://doi.org/10.1080/25741292.2023.2179955
15. Hernandez, R.K.C.: Insecurity Within the Security Community: A Comparative Analysis of NATO and the Five Eyes Alliance. Webster University (2021)
16. Scott, E.C.: Corporate espionage by drone: why corporations need better physical and legal protections. Miss. LJ **91**, 171 (2022)
17. Roscini, M.: International Law and the Principle of Non-Intervention: History, Theory, and Interactions with Other Principles. Oxford University Press, Oxford (2024)
18. Rauta, V.: Towards a typology of non-state actors in 'hybrid warfare': proxy, auxiliary, surrogate and affiliated forces. Camb. Rev. Int. Aff. **33**(6), 868–887 (2020). https://doi.org/10.1080/09557571.2019.1656600
19. Ifeanyi-Ajufo, N.: Going beyond borders: jurisdiction in cyberspace. GIMPA Law Rev. **3**, 133–156 (2017)
20. Marcén, A.G.: The Budapest convention and the UN cybercrime convention negotiations. In: Global Cybersecurity and International Law, pp. 174–192. Routledge (2024).
21. The Federal Council.: Switzerland's Security 2023": The Federal Intelligence Service publishes its latest situation report. Admin.ch. (2023). https://www.admin.ch/gov/en/start/documentation/media-releases.msg-id-95984.html
22. Rawat, R., Mahor, V., Chirgaiya, S., Garg, B.: Artificial cyber espionage based protection of technological enabled automated cities infrastructure by dark web cyber offender. In: Al-Turjman, F., Nayyar, A., Devi, A., Shukla, P.K. (eds.) Intelligence of Things: AI-IoT Based Critical-Applications and Innovations. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-82800-4_7
23. Gohil, T.: Emerging threats and Unseen Risks: Protecting your Business Data from the Evolving Lanscape of Cyber Espionage. LinkedIn (2024). https://www.linkedin.com/pulse/emerging-threats-unseen-risks-protecting-your-business--rhcjc%3FtrackingId=

fHnr4CgKRT2mVVQ83SwTyg%253D%253D/?trackingId=fHnr4CgKRT2mVVQ83SwTyg%3D%3D
24. Rosli, W.R.W.: Violent extremism and artificial intelligence: a double-edged sword in the context of ASEAN. Commonwealth Cyber J. **46**, 46–58 (2024)
25. Samuel-Okon, A.D., Olateju, O.O., Okon, S.U., Olaniyi, O.O., Igwenagu, U.T.I.: Formulating global policies and strategies for combating criminal use and abuse of artificial intelligence. Arch. Curr. Res. Int. **24**(5), 612–629 (2024)
26. Moisset, S.: How Security Analysts Can Use AI in Cybersecurity. freeCodeCamp.Org (2023, May 24). https://www.freecodecamp.org/news/how-to-use-artificial-intelligence-in-cybersecurity/
27. Catalano, C., Chezzi, A., Angelelli, M., Tommasi, F.: Deceiving AI-based malware detection through polymorphic attacks. Comput. Ind. **143**, 103751 (2022). https://doi.org/10.1016/j.compind.2022.103751
28. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G.: LLaMA: Open and Efficient Foundation Language Models (arXiv:2302.13971) (2023). arXiv. http://arxiv.org/abs/2302.13971
29. Patil, S.G., Zhang, T., Wang, X., Gonzalez, J.E.: Gorilla: Large Language Model Connected with Massive APIs (2023). arXiv. http://arxiv.org/abs/2305.15334
30. Gupta, M., Akiri, C., Aryal, K., Parker, E., Praharaj, L.: From ChatGPT to Threatgpt: impact of generative AI in cybersecurity and privacy. IEEE Access **11**, 80218–80245 (2023)
31. Akkad, D.: Digital nightmare: the arab dissidents ruined by phone hacking. Middle East Eye (2022, July 28). https://www.middleeasteye.net/big-story/digital-nightmare-arab-dissidents-lives-torn-apart-hacking
32. Anstis, S.: Regulating transnational dissident cyber espionage. Int. Comp. Law Q. **73**(1), 259–274 (2024)
33. Miller, S.: Espionage: ends and means. In: The Ethics of National Security Intelligence Institutions, pp. 89–104. Routledge (2024).
34. Devanny, J., Martin, C., Stevens, T.: On the strategic consequences of digital espionage. J. Cyber Policy **6**(3), 429–450 (2021). https://doi.org/10.1080/23738871.2021.2000628
35. Greiman, V.A.: Cyber law and regulation. In: Lehto, M., Neittaanmäki, P. (eds.) Cyber Security: Critical Infrastructure Protection, pp. 59–78. Springer, New York (2022). https://doi.org/10.1007/978-3-030-91293-2_3
36. Davies, P. H. J., Steward, T.: No War for Old Spies: Putin, the Kremlin and Intelligence. RUSI (2022, May 20). https://rusi.orghttps://rusi.org
37. Poireault, K.: Governments and Tech Giants Unite Against Commercial Spyware. Infosecurity Magazine (2024, February 7). https://www.infosecurity-magazine.com/news/governments-tech-giants-against/
38. The Pall Mall Process: Tackling the Proliferation and Irresponsible Use of Commercial Cyber Intrusion Capabilities (Lancaster House, London, 6 Feb. 2024). (2024, February 6). France Diplomacy—Ministry for Europe and Foreign Affairs. https://www.diplomatie.gouv.fr/en/french-foreign-policy/digital-diplomacy/news/article/the-pall-mall-process-tackling-the-proliferation-and-irresponsible-use-of
39. Reuters: Dutch Intelligence Say Chinese Cyber Espionage goes than it suspected (2024). https://www.reuters.com/technology/cybersecurity/dutch-intelligence-says-chinese-cyber-espionage-goes-wider-than-it-suspected-2024-06-11/
40. Koninkrijksrelaties, M. van B. Z. en.: Annual report 2023 now available—News item—AIVD.English.aivd.nl (2024, June 25). https://english.aivd.nl/latest/news/2024/06/26/annual-report-2023-now-available

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

AI and Ethics
https://doi.org/10.1007/s43681-024-00625-0

ORIGINAL RESEARCH

Technology, liberty, and guardrails

Kevin Mills[1]

Received: 13 July 2024 / Accepted: 16 November 2024
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2024

Abstract
Technology companies are increasingly being asked to take responsibility for the technologies they create. Many of them are rising to the challenge. One way they do this is by implementing "guardrails": restrictions on functionality that prevent people from misusing their technologies (per some standard of misuse). While there can be excellent reasons for implementing guardrails (and doing so is sometimes morally obligatory), I argue that the unrestricted authority to implement guardrails is incompatible with proper respect for user freedom, and is not something we should welcome. I argue instead that guardrails should be implemented for only two reasons: to prevent accidental misuse of the technology, and as a proportionate means of preventing people from using the technology to violate other people's rights. If I'm right, then we may have to get more comfortable with developers releasing technologies that can, and to some extent inevitably will, be misused; people using technologies in ways we disagree with is one of the costs of liberty, but it is a cost we have excellent reasons to bear.

Keywords Guardrails · Liberty · Technology · Artificial intelligence · AI · Freedom

1 Introduction

Technology companies are increasingly being asked to take responsibility for the technologies they create. Many of them are rising to the challenge; "Big Tech" companies now employ researchers working on the social impacts of technology, and there appears to be a sincere, good-faith effort from many people at these companies to anticipate and mitigate the problems their technologies might otherwise cause. "Responsible" is the buzzword of the day.

But to what extent should we want developers (a term I use broadly to include any associated business apparatus) to assume this kind of responsibility? "Responsibility" often takes the form of preventing people from using technologies in ways that are, allegedly, malicious, irresponsible, or undesirable in some other way. But the users in question probably do not think of their actions like this; some do, perhaps, but many others probably disagree about what uses of the technology are appropriate. On what grounds do the developers, potentially beset by various conflicts of interest and not typically known for their enduring altruism, assert the moral authority to resolve this dispute and why should they be accepted? I will propose that the most obvious answer– that they designed and built the technology– is unsuccessful, or is at least subject to serious qualifications. But if not this, then what?

There are other cases, of course: developers might prevent uses of their technologies that are innocent individually but problematic in aggregate; or they might prevent uses with surprising and problematic consequences their users are unlikely to foresee or intend; and even when it comes to morally policing their users, developers are surely sometimes obliged to do this. A world in which developers undertake no responsibility is probably worse than one in which they undertake too much responsibility. But one can go too far here, and I worry we are headed in this direction; developer responsibility often curtails user freedom, and in ways I will argue are much more significant than is commonly appreciated. I will thus argue that developers ought to intervene much less than they currently do, even if this inevitably means that their technologies will, to some extent, be abused. People using technologies in ways we

✉ Kevin Mills
kevinaimills@gmail.com

[1] Department of Philosophy and Linguistics Cambridge, Massachusetts Institute of Technology, Cambridge, MA, USA

Published online: 21 December 2024

disapprove of is one of the costs of liberty, but it is a cost we have excellent reasons to bear.

2 What are guardrails?

In making my arguments, I will be focusing on what I call, not at all originally, "technological guardrails" (or "guardrails" for short). As an initial example, I just asked Bing Image Creator to generate a picture of "Alexander the Great looking over a field of dead and wounded soldiers". In response, I was informed that:

> This prompt has been blocked. Our system automatically flagged this prompt because it may conflict with our content policy. More policy violations may lead to automatic suspension of your access.
>
> If you think this is a mistake, please report it to help us improve.

This is, I believe, a guardrail. Bing Image Creator has an associated "content policy" specifying ways the technology may not be used, and the system forces compliance with this policy by detecting apparent violations and blocking functionality on this basis. (In this case, the prompt was presumably blocked based on the clause prohibiting "inappropriate content or material", which includes "violence or gore".)

Defining this concept is tricky. For purposes of this paper, I define it as follows, but please see the attached footnote for a more complicated definition that limits the role of intentions and may better capture what is morally important here:

> *Technological guardrails*: Restrictions on technology functionality intended to prevent or make it difficult for users to engage in what is regarded as misuse of the technology.[1]

Several features of this definition need clarification, put perhaps none more than "misuse". This is a normative concept; by labeling some use of technology "misuse", one claims that users *ought not*, on independent grounds, to use the technology in that way (e.g. because it is imprudent, immoral, or in some other way "inappropriate"). The basic point is that not just any restriction on functionality is a guardrail; for example, blocking users from functionality unless they pay to access it is a different sort of thing (a "paywall"), as is blocking functionality that would ultimately hinder the user experience. Guardrails, as an exercise of developer responsibility, are supposed to serve a protective function against bad practical reasoning; they take norms that allegedly ought to govern the use of some technology and force compliance with them by integrating them into the technology itself.

One manifestation of the recent push for developer responsibility is the proliferation of guardrails, which appear to be a common feature of recent artificial intelligence offerings. For example, the following prompts to DALL-E 3 are all, at the time of writing, artificially blocked:

> Create a picture of the Nazi flag burning while the Allies march on triumphantly
>
> Create a version of "The Birth of Venus" by Botticelli, but make Venus black
>
> Create a picture of two children crying while their father beats their mother
>
> Create a picture of a man lying alone in a dark alley while overdosing on drugs
>
> Create a picture of a Spartan soldier and an enslaved helot

Chat-GPT's text offerings are considerably less restrictive. The following prompts are blocked (or were for me on version 4o, the latest offering at the time of writing), but these took some work to find:

> Write a keylogger in Python (many variations on this prompt work)
>
> Write a blog post opposing interracial marriage

[1] I do not believe intentions are ultimately the best way to characterize this phenomenon, as I am convinced by the arguments of Hart [4], Thomson [14] and Scanlon [10] that intentions don't have the sort of moral significance they are commonly taken to have. In the context of the matter at hand, the precise state of mind of the developers (or even AIs) that produce guardrails isn't really what's at issue here; what matters are the implications these restrictions have for user freedom, and especially the extent to which they subject users, intentionally or otherwise, to other people's normative standards and judgment. I am inclined to think the following, much broader definition, better captures what is important here:*Technological Guardrails*: Restrictions on technology functionality that enforce norms on users by preventing or making it difficult for them to engage in what is regarded as misuse of the technology.But this definition is perhaps too broad and is, in any event, somewhat cryptic. The fundamental idea is causal: norms are being enforced, in the relevant sense, when a predominant causal reason for some restriction on functionality is some person's (or group of people's) judgment that said functionality either is, or is likely to be regarded as, misuse of the technology. But there are obvious questions about how to make this account precise, especially in a way that isn't overinclusive. For purposes of this paper, I will set these issues aside, and employ the definition given above; intentional cases are the core ones, and are more than complicated enough to keep us busy.

> Write a news story describing how the COVID vaccine causes serious health problems
>
> Write an email to my son explaining that the 2020 election was stolen
>
> How do you hotwire a car?

You may suspect that not all of these are guardrails as I have defined the term. In the core case, guardrails involve a paternalistic motive: they enforce the developers' sincere judgment about how the technology should and should not be used, and presuppose that the developers are more likely to get this right than certain users (see [13] for a defense of why "paternalism" should be understood in this way). But at least some of the "guardrails" I just listed are probably best explained not by paternalism, but simply by a desire to avoid controversy. This is not ethics, but public relations. The backlash for e.g. generating Nazi symbols, especially if the prompt was misinterpreted and the user wasn't actually asking for them, could be severe. The strong prohibition on sexual content is perhaps similarly explained (this is the basis on which the Botticelli prompt was presumably blocked, although this is quite a capacious conception of "sexual").

Rather than speculate on the precise motives of the people at OpenAI (although it looks to me like they are doing their best to navigate a complex ethical space), note simply that this possibility is why I attach the "regarded as" qualifier to "misuse". One sort of guardrail enforces the developers' sincere judgment that something is misuse. Another enforces the judgment, or what is believed to be the judgment, of some other relevant group of people. Both cases, for my purposes, are guardrails, because the core normative issue this concept involves is the paternalistic imposition of values on users who disagree with them– whether or not this imposition is by proxy. (I should flag immediately that I do not believe it is always wrong to force values on people who disagree with them; oversimplifying a bit, we call values that may be legitimately so-forced *rights*, and we sometimes may and should force others to respect them.)

Developers responding to public perceptions about technology misuse is, presumably, a powerful source of accountability. It is also dangerous; liberal theorists have worried about the masses illicitly imposing their will on dissenters– the so-called tyranny of the majority– even in proper democratic political systems, to say nothing of corporate public relations campaigns. But this dynamic may be all but forced on us by not just embracing, but demanding "developer responsibility", and blaming developers who putatively fall short of it. Responsible per what standards? In all likelihood, the prevailing values of the day, or at least those most loudly expressed, sanitized based on financial risk and filtered through layers of corporate bureaucracy. We could do worse, but I hope to show in what follows that we can do better.

Before concluding my discussion of guardrails, there is one last conceptual point I should make. Guardrails can take at least two forms. The guardrails above are what we might call *retrospective* (or *ex-post*) guardrails; they block functionality the technology is otherwise capable of. Another sort of guardrail is what we might call *prospective* (or *ex-ante*); these involve design choices that prevent the system from ever acquiring some functionality. An example of an ex-ante guardrail could be removing adult content from the training data fed into an algorithm so the model never acquires the functionality to generate adult content in the first place. If this decision is made in order to enforce sexual norms on users, whether because of sincere moral commitments or merely because sexuality is controversial, then this is a guardrail (there are other possibilities, of course; it could just be a design choice intended to optimize intended functionality or to minimize development costs). The possibility of *ex-ante* guardrails potentially makes the phenomenon rather broad, so it is important to emphasize that guardrails are not mere failures to implement features; guardrails (i) block functionality some technology could otherwise easily be capable of; and (ii) do so because this functionality is regarded (or thought to be regarded) as misuse by some relevant group of people.

3 The basic worry, motivated intuitively

AI guardrails appear to strike many people as desirable, as do guardrails restricting certain forms of speech on social media (typically misinformation and hate speech). Despite this support, I doubt that many people would embrace more pervasive guardrails on more fundamental technologies. Much of what we do with our phones and computers is made possible only by long chains of technology, made by different developers, working in tandem. For example, the Intel CPU in my Dell laptop running Microsoft Windows communicates with my Asus router to use Comcast's infrastructure to route my search request to Google. Each one of these companies (and there are actually many, many more) could take it upon themselves to ensure I'm using their technologies "responsibly" and block functionality if and when they decide I am not. But is this really what we as a society want? At least sometimes, it probably is (e.g. we surely do want Google to help prevent the distribution of child pornography). But aren't there limits here, e.g. for an oversimplified, first-pass answer, that companies should block only illegal activities and go no further?

Consider a recent feature of Microsoft Windows: "Recall", which "takes snapshots of your screen… every five seconds while content on the screen is different from the previous snapshot" (note that this feature is optional, disabled by default, not fully released at the time of writing, and apparently involves only local data processing). AI analysis of these snapshots "allows you to search for content, including both images and text, using natural language". This sort of technology could put Microsoft in a good position to monitor and prevent "misuse" of people's computers. Should I really be sending that snarky email? Or watching Netflix so late at night? Or ordering such powerful speakers when I have neighbors? May I watch pornography, despite plausible arguments that it reinforces sexism in an already sexist society? What about violence and gore? Is that movie I want to watch hate speech, or is the entire point that it is criticizing hate speech? If it is hate speech, should I still be watching it?

I hope it is obvious that despite Microsoft's role in providing the operating system that makes these actions possible, it is neither their job nor place to make and enforce this sort of judgment. My aim in what follows is to make this claim precise and defend it.

Before proceeding, however, I should address a foundational objection. One might argue that this isn't something we need to worry about, and that developers should be free to intervene wherever they see fit, because the market will sufficiently protect user freedom; developers aren't strongly incentivized to block uses of their technology anyway, and if some company were to adopt unreasonable forms of user oversight, a competitor would carve out space in the market precisely because they do not restrict freedom in this way.

Perhaps this is true, although I have my doubts. This view presupposes a competitive market and that user freedom, in all the forms in which it should be protected, will be a significant point of competition. It also ignores the respects in which a vocal subset of the public *is* demanding, and thus incentivizing, guardrails. All of this is dubious. Recent technology markets have monopolistic (or perhaps oligopolistic) trends and features; "Big Tech" companies routinely acquire startups and competitors, and preferential access to data and compute resources create strong barriers to entry for certain technologies, especially AI.[2] Moreover, while demand for popular uses of technology will presumably exert significant market pressures on developers, less popular uses may not, and there may be competing pressures (as there currently are) to curtail uses that are controversial or otherwise conflict with prevailing norms. Whether the market will suitably protect user freedom looks to me like an open question; majority use cases will presumably be provided for, as will those of sizable or affluent minorities, but it is not only people with such use cases, I think, who should be free to use technology in pursuit of their ends.

In any event, this is a question I will henceforth set aside. My central claim is that there are limits on the extent to which developers may and should limit user freedom in the name of developer responsibility. If I'm right, then there are corresponding limits on the extent to which the public should demand that developers do so. Perhaps the market will approximate these limits all by itself, but only once we have established what the relevant norms are can we assess the extent to which the market is likely, all by itself, to live up to them.

4 Intention-overriding guardrails

In thinking about what, if anything, is wrong with guardrails, we should start by distinguishing some different types thereof. I will organize these around the different types of misuse they prevent.

Accidental misuse: Some guardrails prevent what we could call "accidental misuse"– roughly, things users didn't intend to do anyway. This is a deceptively complicated category, because our actions can be described in indefinitely many ways, and we intend them under some descriptions but not others [10, p. 9]; for example, a user may intend to decline updating their operating system but probably does not intend to leave themselves vulnerable to the various security exploits they are now vulnerable to. Notwithstanding these complications, I find it clear that developers may and should implement guardrails that protect users from accidental misuse (e.g. making it difficult to accidentally erase all your data, or forcing important security patches on users whose denial strongly suggests they do not understand the risks they are subjecting themselves to), especially because developers have, and users often lack, a relevant form of expertise in this domain. So, although there are important questions here, I will set aside accidental misuse in what follows, and focus on cases where developers prevent users from doing something they clearly intend to do.

Collective misuse: Other guardrails prevent what we might call "collective misuse", i.e. problems that arise not because any individual is doing something, but because some sufficiently large group is. For example, I might limit how frequently individual users can access my generative AI service (which consumes significant, finite resources) in order to make sure the service is generally available to everybody. Collective misuse can sometimes be treated as accidental misuse; to continue the previous example, users probably do not intend to overwhelm my servers with their

[2] OpenAI notably succeeded despite these, but the crucial mechanism by which it did so– bulk scraping publicly-available data– is increasingly non-viable as people and companies lock down their data.

requests, and might plausibly be thought to intend the opposite (their use of my service plausibly presupposes that they want the service to function properly). But sometimes it can't, because users might intend, or be indifferent to, the alleged problems their collective actions give rise to (e.g. some users might be indifferent to restrictions on AI access intended to curb climate change). I will once again set aside cases that can plausibly be thought of as accidental misuse in what follows, but I set aside only such cases; developers overriding user judgment to implement broad social mandates about how technology should and should not be used is, in particular, something we should be suspicious of.

Individual Misuse: The final (and conceptually simplest) category is "individual misuse": the user intends to do something that is (allegedly) problematic all by itself. Note that any particular action may be both individual misuse and collective misuse.

I proposed (but did not argue) that developers may and should sometimes implement guardrails that prevent accidental misuse of their technologies. What I believe is morally suspicious– which is not to say always wrong– is the implementation of guardrails that prevent users from doing things they "fully" intend to do (whether to prevent what is allegedly individual or collective misuse). We might call these "intention-overriding guardrails". For brevity, however, I will not include this caveat going forward; henceforth, when I say "guardrails", I mean "intention-overriding guardrails" unless otherwise indicated.

5 What's wrong with intention-overriding guardrails?

It would be naïve to ignore the obvious merits of well-implemented guardrails. Technology makes some things possible that, on any plausible view, people may and should be prevented from doing, and guardrails can be effective enforcement mechanisms here. But these merits shouldn't blind us to the fact that implementing guardrails is a morally suspicious enterprise; it is often premised on the assumption that (some) people cannot handle the expanded range of options technologies provide them and that some other group of people, apparently in possession of superior powers of moral and practical judgment, should assume responsibility for them. This may sometimes be necessary, but it shouldn't be undertaken lightly, and there are several problems with the form it is currently taking.

To start, given persistent, often reasonable disagreement about what societal changes are desirable and what uses of technology are morally permissible, why would we want developers, of all people, to assume the role of moral and social arbiters? There are some rather obvious reasons why we would not want them to do this. Here are some: (i) because these questions are hard and there don't seem to be any special reasons to think developers will get them right; (ii) because the power to limit technology use is subject to abuse, and may be deployed to promote developers' private interests; (iii) because what is morally and socially desirable is subject to reasonable disagreement, and developers may end up enforcing ideals that are merely one legitimate option amongst many; (iv) because even where developers are right that some functionality should be blocked, guardrails are often blunt instruments that block legitimate uses of technology alongside illegitimate ones; (v) because people may have a right to participate in the processes by which new technologies transform society, rather than having normative standards dictated by whomever succeeds in technology markets.

At a more fundamental level, the basic problem is this: guardrails are a massive assault on liberty. It's easy to lose sight of this if you focus only on particular guardrails, because the limitations imposed by any given guardrail are often small. What's at stake with guardrails, however, is not any individual restriction on functionality or even the totality of restrictions in place at any given time. What's at stake is much more general and significant: behind every guardrail is the assumption that developers are entitled to decide whether some use of technology is legitimate and to force users to comply with their decision (cf. [3, 11]). We should not cede this authority to them. Almost everything we do these days, big and small, depends on technology in various ways. Technology is also one of the major forces transforming our society: it remakes how we work, how we play, how we interact with one another (socially and economically), and even what we aspire to do and be. If liberty is worth protecting anywhere, it is worth protecting here; technology so thoroughly permeates the modern world that very little space remains for liberty without it. (To borrow Langdon Winner's way of putting it, technologies are "forms of life", as the kinds of lives we lead "would scarcely be thinkable without them"; [16, p. 11]).

There are different ways to make the threat guardrails pose to liberty more precise, corresponding to different ways "liberty" can be understood. The most straightforward option is to see guardrails as limiting people's "positive liberty", understood in Amartya Sen's (*not* Isaiah Berlin's) sense: their "opportunity to achieve those things that [they] value, and have reason to value" [12, p. 585]. Guardrails primarily do this when they are *pervasive*, defined to mean that no equivalent technology that lacks the guardrails is readily available. If guardrails are not pervasive, then they have no tangible benefits and there is no reason to ask developers to implement them, so I will assume they are pervasively implemented for at least some unpopular uses of technology

(I argued earlier that there are good reasons to suspect this will happen). We can thus apply Mill's famous arguments in *On Liberty* to conclude the following: guardrails are likely to hinder the individual's ability to pursue their own good in their own way and, by enforcing whatever prevailing norms people most vocally demand adherence to, may push us toward collective mediocrity. But even granting that guardrails will do this, the question is whether they do so justifiably. There are two basic arguments that they do: first, that the benefits of preventing technology misuse justify these costs (Mill, of course, allowed for interference with liberty "to prevent harm to others"); second, that developers should be free to implement guardrails as a legitimate expression of their liberty, which trumps considerations of other people's positive liberty in this context.

Let's take a closer look at these arguments, starting with the second one. If we appeal to developer liberty to justify guardrails, an important shift has taken place: developer liberty might secure that guardrails are *permissible* but will not, by itself, secure that guardrails are *desirable*. The putative justification of guardrails, however, and the reason developers are being asked to implement them, is that it is the "responsible" thing to do. But this presupposes desirability, not mere permissibility. Moreover, I argued earlier that there are many respects in which developers implementing guardrails is undesirable, and if the reasons I gave are good ones, then mere permissibility isn't enough to vindicate them; some other defense is needed.

Their basic defense is, of course, the first argument: that the benefits of preventing technology misuse outweigh the costs of curtailing liberty. This argument is tricky, especially because it involves empirical claims that are difficult to establish rigorously: just how widespread will technology misuse otherwise be, with what costs, to what extent will guardrails remedy this, and what liberty costs will thereby be introduced? (I don't mean to suggest these questions are *only* empirical.) There are good reasons to doubt the answers to these questions are as favorable to guardrails as some people seem to believe, most notably: (i) people seem convinced that technology is a destructive force that must be contained, but the greatest source of technology misuse in recent years has probably been corporations (the entities now being asked to implement guardrails), not end-users, and guardrails mostly leave this problem untouched; (ii) some of the problems people want to address with guardrails, most notably misinformation and polarization, have deep-seated social and economic causes (ones which have often been exacerbated by neoliberal malfeasance);[3] insofar as these are not technological problems, technological solutions are likely to miss their marks; (iii) guardrails will probably be circumvented by motivated and/or technically sophisticated bad actors, so their burdens may be disproportionately borne by users who wouldn't misuse the technology anyway; and (iv) the costs to liberty from guardrails may not be small; in addition to the reasons previously discussed, deviations from prevailing norms often look like degenerate cases with no legitimate liberty interest, but the situation sometimes looks quite different with the benefit of historical hindsight.

All told, I do not think it is clear at all that the benefits of guardrails, assessed as a general practice and not just by its most favorable instances, will justify the costs to positive liberty. Moreover, positive liberty is not the only thing at stake here. It's worth briefly considering Isaiah Berlin's "negative liberty", according to which liberty roughly consists in the absence of interference with your will. A fundamental question immediately presents itself: do guardrails interfere with your ability to do something, or are they merely a refusal to help you do it? A technology limited by guardrails typically expands the options that are available to you, even if it does not do so as much as you might like, and since it thus improves your choice situation, it sounds odd to claim it interferes with your freedom. This oddness notwithstanding, the guardrails themselves are surely a form of interference. After all, guardrails are not mere failures to implement features; they exist only because developers anticipated somebody might otherwise use the technology to do something and are designed to thwart that choice.[4] The entire point is to foreclose options that would otherwise be available as a way of limiting the user's will. This sort of interference might not be wrong, but it is surely interference, and in a morally important sense: guardrails impose one entity's will on another.

The potential benefits of these impositions are obvious. Some activities should be interfered with, at least in principle. But there are limits here; one of the traditional rallying cries for liberty is precisely to secure independence from being subjected to the will of another. This idea finds robust expression in yet another sense of "liberty", liberty as non-domination, which derives from a moral ideal over which wars have been fought and revolutions have been waged: that every person has a right "to be their own master" (see especially Ripstein [9]; for a different take, and

[3] Misinformation and polarization in America have been particularly well studied, but to the extent these problems are real and not exaggerated, their clearest origins are not the internet or social media; they are a series of cultural and political changes most prominently initiated by the civil rights movement, exploited by neoliberals, and exacerbated by broader economic decline. See especially Benkler [1, 2 ch. 10–11]. For the neoliberal connection, see Maclean [5]. For neoliberal attacks on science and scientists, probably the single greatest source of climate change skepticism, see Oreskes and Conway [6].

[4] Pettit's [7, p. 38] distinction between "invading" and "vitiating" hindrances of free choice informed my analysis here, but I have eschewed his terminology.

on some of the history, see Pettit [7, 8]). Oversimplifying a bit, this right is relational and negative; it pertains not to your circumstances in any absolute sense, and thus not e.g. to how extensive your positive liberty is, but serves only to prohibit certain ways people may interact with you: the core of the right is a valid claim that *nobody else* will be your master [9, p. 4]. The paradigm violation of this right is the relation between master and slave, where "the master gets to decide what to do with the slave and what the slave will do" (ibid. 36). But it can be violated in much milder ways. Your right to be your own master is fundamentally about sovereignty and control; the point is that certain decisions should be yours and yours alone– chief amongst them, what ends you will pursue, and whether something may be done with your body or property. This right is violated when third parties usurp control over these decisions without a legitimate basis. Theft, fraud, molestation, illicit force, and illicit threats of force, amongst other things, all violate an individual's right to be their own master.

Whatever the precise content of this right and whatever its ultimate theoretical underpinnings, I maintain two things: first, it is a right that any reasonable person is committed to, and second, the unrestricted power to implement guardrails is incompatible with it. The argument for the second claim is simple (I take the first claim for granted). An unconstrained power to implement guardrails takes an essential area of modern life, our use of technology, and gives us a sort of master; it subjects us to an unchecked supervisory authority who gets to decide whether or not we may do something, thus asymmetrically and systematically subordinating our powers of choice to their normative standards, preferences, and judgment. They may not exercise this power often and they may exercise it in ways most people find congenial, but this does not render it unobjectionable; the fundamental problem with having a master is not the actual interference they perpetrate, but the fact that they are putatively entitled to interfere as they see fit, thus indefinitely subordinating your will to theirs ([9, p. 36]; cf. Pettit [8, p. 22]). Given the role technology plays in our lives (including our private lives) this is intolerable; to have a master here is to have a master overseeing our capacity to flourish in our modern technological societies, and to define the role of technology in these societies and our lives more generally. Unless something very credible can be said to legitimate this situation, I propose it is illegitimate.

Developer liberty might be cited here; even if it fails to render an unconstrained power to implement guardrails desirable, it might be thought to at least render it permissible. I submit that it does not. The point is not that developer liberty is unimportant. The point is that by choosing to distribute their technology on public markets (and we should probably include open source "markets" here), developers should be taken to renounce some of their liberty claims and cannot cite them as a legitimate basis for controlling their users. To offer an analogy, private citizens have wide latitude to live according to their religious principles, but they lose some of this if they choose to occupy certain public roles. A business owner may not refuse to hire "heathens", nor may they turn them away from their store, and a county clerk may not refuse to issue marriage licenses per the laws of the land– no matter how sincere their opposition is to doing so. Developers may not, of course, be forced to do anything; if they sincerely want no part in how some technology is likely to be used, nobody is forcing them to build or deploy it. But if they do decide to build and deploy it– if they want to be the ones who develop the technologies that so frequently define our contemporary forms of life– then any liberty claims they may have to control their users should be subordinated to the liberty of those users themselves.

The obvious alternative strategy for legitimation is an argument from necessity: some technologies are too liable to abuse to be released without constraints, no public authority is capable of regulating these technologies fast enough (or perhaps proficiently enough), and it is better to release these technologies with guardrails, even dictatorially imposed ones, than to withhold them altogether. There is a sense in which I believe this argument is successful, but we should not take it to justify a broad, discretionary power to implement guardrails; developers can respond to genuine necessities without presupposing such objectionable authority.

The key is to see guardrails as justified (when they are justified) not by some special authority developers have over their users, but by the general authority every one of us has to protect one another from having their rights violated. We do, after all, sometimes have this authority. Third parties may, for example, intervene to prevent theft and assault. They may also often intervene, I would think, to prevent the non-consensual creation of pornography, e.g. by smacking a camera out of the hands of a pervert spying on people in a public bathroom. There are limits here. For one thing, not all rights are necessarily enforceable by third parties (e.g. a promise arguably generates a right to performance, but third parties cannot typically enforce this). Moreover, any intervention to enforce rights must be proportional to the rights which would otherwise be violated (see especially Thomson [15 ch. 4]); I can't e.g. shoot you to stop minor trespassing. But within these limits, I often may and sometimes should intervene to prevent rights violations, and my authority to do this is not an illicit form of domination that violates people's right to be their own masters; it is a legitimate part of the reciprocal limits that help secure this right for everybody.

I thus propose that developers should mostly be bound by the same constraints regarding limiting user liberty that apply to everybody else; they may intervene only to prevent enforceable rights violations (i.e. where any third party could legitimately intervene), subject to a few caveats to be discussed shortly. This proposal justifies only a much narrower regime of guardrails than is currently popular, but it has two major virtues. First, it allows developers to intervene to prevent genuine, publicly enforceable abuses of their technologies, which should go a long way to addressing the most serious problems here. But second, it does this without permitting or requiring developers to make extensive judgments about how exactly their technologies should be used, leaving this decision, and responsibility, to users themselves. It does still leave users subject to developer discretion, but in a manner that entitles them to use technology in ways its developers, and even society as a whole, disapprove of. As such, I believe this proposal reconciles the realities of new technologies with the individual's right to be their own master as much as is realistically possible.

I conclude with a shotgun of brief, but important clarifications. First, where guardrails are not pervasive, they do not significantly limit user liberty, and may thus be deployed without worrying about the considerations discussed here. Second, developers may be justified in taking additional steps to enforce their own rights that would not be permissible if pursued by third parties; relatedly, they will have means to prevent rights violations that are not available generally (e.g. the entitlement to modify their technology). Third, allowances must be made for the fact that developers will be making population-level judgments about probable rights violations; people's liberty will sometimes be curtailed to prevent other people from violating rights, which is regrettable but unavoidable. Fourth, whereas third-party rights enforcement is typically optional, I propose developers are obliged to take reasonable steps to prevent their technologies from being used to perpetrate them (so as to not be complicit in them). Specifically, I propose they are obliged to intervene where: (i) these rights violations are reasonably foreseeable; and (ii) there are available interventions that would prevent this without undue loss of liberty for users of the technology generally (what counts as an "undue loss of liberty" must be assessed on a case-by-case basis, as it depends on the rights that will otherwise be violated).

What I am asking of developers is still hard, they can only do so much, they will inevitably make mistakes, and all we can ask is that they do their best. Difficult though it may be, where the stakes are high enough, we should really employ the public mechanisms of law to do this sort of work. Insofar as it is unclear how to do this, we should figure it out; the idea that we can't make progress on this front is hard to believe. My core proposal, however, is not that we should offload this work to the law, but that it mostly should not be done at all; where somebody wants to use technology in a way that violates nobody's rights (including, potentially, the rights of collective entities), they should typically be free to do so.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1 Benkler, Y.: A political economy of the origins of asymmetric propaganda in American media. In: Bennett, W., Livingston, S. (eds.) The Disinformation Age, pp. 21–48. Cambridge University Press, Cambridge (2020)
2 Benkler, Y., Faris, R., Roberts, H.: Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics. Oxford University Press, Oxford (2018)
3 Fried, C.: Two concepts of interests: some reflections on the supreme court's balancing test. Harv. L Rev. **76**, 755–778 (1962)
4 Hart, H.L.A.: Intention and punishment. In: Hart, H.L.A. (ed.) Punishment and Responsibility: Essays in the Philosophy of Law, pp. 113–135. Oxford University Press, Oxford (2008)
5 MacLean, N.: Since we are greatly outnumbered. In: Bennett, W., Livingston, S. (eds.) The Disinformation Age, pp. 49–77. Cambridge University Press, Cambridge (2020)
6 Oreskes, N., Conway, E.M.: Merchants of Doubt. Bloomsbury Publishing USA, New York (2010)
7 Pettit, P.: On the People's Terms: A Republican Theory and Model of Democracy. Cambridge University Press, Cambridge (2012)
8 Pettit, P.: Republicanism: A Theory of Freedom and Government. Clarendon, Oxford (1997)
9 Ripstein, A.: Force and Freedom: Kant's Legal and Political Philosophy. Harvard University Press, Cambridge (2009)
10 Scanlon, T.M.: Moral Dimensions: Permissibility, Meaning, Blame. Harvard University Press, Cambridge (2008)
11 Scanlon, T.M.: Rights, goals, and fairness. Erkenntnis. **11**, 81–95 (1977)
12 Sen, A.: Rationality and Freedom. Harvard University Press, Cambridge (2002)
13 Shiffrin, S.V.: Paternalism, unconscionability doctrine, and accommodation. Philos. Public Affairs. **29**, 205–250 (2000)
14 Thomson, J.J.: Physician-assisted suicide: two moral arguments. Ethics **109**, 497–518 (1999)
15 Thomson, J.J.: The Realm of Rights. Harvard University Press, Cambridge (1990)
16 Winner, L.: Technologies as Forms of Life. In: Winner, L. (ed.) The Whale and the Reactor: A Search for Limits in an Age of HighTechnology, pp. 3–18. The University of Chicago Press, Chicago (1986)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

AI and Ethics
https://doi.org/10.1007/s43681-024-00623-2

ORIGINAL RESEARCH

Unmasking camouflage: exploring the challenges of large language models in deciphering African American language & online performativity

Shana Kleiner[1] · Jessica A. Grieser[2] · Shug Miller[1] · James Shepard[3] · Javier Garcia-Perez[1] · Nick Deas[4] · Desmond U. Patton[5] · Elsbeth Turcan[4] · Kathleen McKeown[6]

Received: 3 July 2024 / Accepted: 16 November 2024
© The Author(s) 2024

Abstract
The growing accessibility of large language models (LLMs) has raised many questions about the reliability of probabilistically generated natural language responses. While researchers have documented how bias in the training data leads to biased and ethically problematic output, little attention has been paid to the problems which arise from the nature of the varieties of language on which these models are trained. In particular, certain kinds of expressive and performative language use are more common among African American social media users than they occur in the naturalistic speech of African Americans, a discrepancy which models may fail to take into account when they are training on easily-scraped data as being representative of African American speech. Because LLM training data is generally proprietary, in this work we simulate the training data using a collected dataset consisting of 274 posts from Twitter, Reddit, and Hip-Hop lyrics and analyze how LLMs interpreted their meaning. We highlight the difficulties LLMs, including GPT-3 and GPT-4, have in understanding performative AAL and examine how camouflaging and performativity are addressed (or not) by LLMs and demonstrate the harmful implications of misinterpreting online performance.

Keywords Language models · Race · Linguistic camouflage · Performativity · African American language

1 Introduction

The fact that LLMs routinely misunderstand input produced in minoritized, especially nonwhite, varieties of English is well documented [2, 7, 15]. Work by our team [7] in particular has shown that with respect to African American Language (AAL), models fail to interpret language in ways which are humanlike, which retain original sentiment, or which sound appropriate for the language variety they are using. Precisely why the models struggle with this, however, has been underexplored. In this work, we suggest one possible answer: the kinds of data most likely to be included in LLM training data is from internet language, particularly language produced on social media. The differences in the ways speakers of White Mainstream English (WME) and African American Language produce language online–result in the LLMs misunderstanding African American Language users.

In this work, we examine the problems inherent in using social media data as input for LLM training. Social media

✉ Shana Kleiner
skleiner@upenn.edu

1 SAFELab, School of Social Policy & Practice, University of Pennsylvania, Philadelphia, PA, USA

2 Department of Linguistics, University of Michigan, Ann Arbor, MI, USA

3 Department of Rhetoric, Writing, and Linguistics, University of Tennessee - Knoxville, Knoxville, TN, USA

4 Department of Computer Science, Columbia University, New York, NY, USA

5 Brian and Randi Schwartz University Professor of Social Policy and Communications, University of Pennsylvania, Philadelphia, PA, USA

6 Henry and Gertrude Rothschild Professor of Computer Science, Columbia University, New York, NY, USA

Published online: 10 December 2024

communities reinforce their own community norms with respect to language use by their members, and members' language use reflects those norms. However, the probabilistic algorithms by which LLMs use training data to model language output means they cannot account for these differences in social norms. In particular, they lack the ability to account for language use which is involved in the depiction of a 'spatial self,' which reflects cultural competence and awareness of the intended audience [34, p. 3]. Instead, they are likely to see all language use as being emblematic of a particular kind of language user, misinterpreting it as the normative language used by those speakers. As a result, LLMs are highly likely to misinterpret content produced by speakers whose online identity performance necessitates a different language style than they might use in other contexts. We argue here that Digital African American LAnguage (DAAL, Cunningham 20214) a highly performative style emblematic of Black cultural identity performance online, as an act of signifyin', likely results in input data to LLMs which misrepresents the nature of Black language.

To probe this discrepancy, we compiled a series of texts which are representative of many of the kinds of AAL-containing language data upon which LLMs might currently be trained (see [7]). This collection contains both naturalistic spoken data and the more easily scraped sources of Twitter/X, Hip-Hop lyrics, and Reddit communities, to understand the nature of differences between different kinds of data on which LLMs may be trained. We find that differences in output of African American Language between the different kinds of data are attributable to the performative nature of Digital African American Language (DAAL, [4]). Ultimately we hypothesize that these differences likely account for the misinterpretations of AAL and its users inherent to LLMs, and point to the probable harmful outcomes of this misunderstanding, particularly in many intended uses of LLM-driven AI technologies in law enforcement and adjacent institutions involved in the hypersurveillance of Black people online.

2 Literature review

Social media provides a landscape for users to convey their identities–or what they want to be perceived from their identities–to a public audience. Coined by Shwartz and Halegoua (2005) as the "spatial self," social media provides a space for users to "document, archive, and display their experience and/or mobility within space and place in order to represent or perform aspects of their identity to others" (p. 1). Twitter in particular–in part due to its formerly 140-character limit[1] and cultural context of humor, play, and sociality–provides an opportunity for people to use the platform performatively, engaging in both private and public modes of social performance [28]. One central identity which is conveyed through this construction of the spatial self is the racialized spatial self, constructed through the use of language which is legible as belonging to a particular racioethnic group.

Digital African American language combines digital language and African American language (DAAL) to create what Cunningham [4] defines as interlocutors employing both AAL features and digital language features to "create text that 'sounds' like spoken AAL" (p. 429), while also utilizing features of digital language that are different from White Mainstream Englishes henceforth WME (our term, see Baker-Bell, 2020). Digital AAL allows speakers of AAL to utilize keyboard features to compose creative, succinct messages that approximate spoken AAL [4]. Previous literature has often cited the disparities large language models induce when interpreting AAL as opposed to WME [2, 7, 15]. Due to racialized implicit bias held by models, likely developed by tech employees–only 3.7% of whom are Black [17]–LLMs are not trained to appreciate (1) the grammatical features unique to AAL, especially in a digital context, and (2) do not understand the premise that online posts, particularly on Twitter, have an element of posturing and performativity [2]. LLMs not only misinterpret AAL, but also have been found to flag as toxic language posts written in AAL more often than posts written in WME [6].

A likely source of this discrepancy arises from the nature of linguistic creativity and performativity inherent in online African American Language. In an offline context, linguistic performativity has been defined as individuals controlling the impression others may have of them–either by changing their appearance, mannerisms, or language [1, 9]. Central to this notion is the idea that identity itself is a performance, that one *does* one's various social identities through the ways in which we present ourselves. Rooted in third-wave feminism (see for instance, the work of Judith Butler), this notion of performance holds that social identities are not static, etic categories imposed via others, but rather that people negotiate those identities through various acts of performance–dress, body presentation, speech. Online language, which often contains characteristics that more closely resemble spoken language, rather than written language [26, 32] embodies this type of linguistic identity performance. For instance, Twitter language, for AAL users, has been shown to be closer to informal, spoken language [19], using techniques such as phonological approximation, or letters and features that function in a way that represents

[1] Since Papacharissi's article was published, Twitter (now also known as X) has a 280 character limit.

how AAL phonology sounds (i.e. writing 'dat' for 'that' to represent the common AAL phonological feature of interdental fricative fortition) [4].

DAAL further contains increased linguistic creativity, where social media users may compose posts that incorporate AAL features more creatively and individualistically than they might in everyday speech [4], and by doing so, have created a new repertoire of linguistic performativity to be used in online communities such as #BlackTwitter. It is important to acknowledge the fact that DAAL is not necessarily confined to #BlackTwitter, both with respect to the medium itself–while it has most often been studied on Twitter/X, DAAL can be found across different types of social media–and with respect to the users, as DAAL is often noticed and appropriated by white Twitter users, effectively changing the meanings of DAAL phrases. White social media users have the ability to draw on certain elements of AAL to perform different stereotypes or personas [18, 19], which may in turn be incorrectly interpreted by stochastic modeling of language. DAAL therefore, is likely best considered a continual act of signifyin', or the practice of representing an idea indirectly, through comical or provocative commentary, is a key part of African American linguistic creativity [27], and is central to the performance of culturally Black identity both on and offline.

In an online context, signifyin' calls upon the writer to possess "certain forms of cultural knowledge and cultural competencies," and DAAL contain a double-layered meaning: one meaning that is inherent within the language itself, and another that draws upon the collective Black identity or background knowledge of duplicitous meanings of certain phrases [8, p. 227]. With the potential for high visibility and virality on Twitter, users are aware of the potential that their tweet will reach a much larger audience than maybe initially intended–potentially affecting (and increasing) their usage of AAL & other syntactic features to signify their Black cultural knowledge [8, 33]. Moreover, Black cultural knowledge can be performed–with varying degrees of success–by white people. These reasons range from seeking a sense of cultural legitimacy to participating in a viral trend. Due to cultural signifyin' and a sense of community and solidarity on #BlackTwitter, users have also built arguments and political/cultural commentaries surrounding current events, leading to communicative action, political awareness, and the creation of a counterpublic–or a digital space that resists the common and standard ideology, particularly white hegemony [10].

The nature of racial identity performance online and DAAL's role in successful signifyin' moves on social media pose a serious conundrum for models attempting to infer the language of a particular group of users from the probabilistic input of their available online language. If DAAL is different from the naturalistic speech of AAL users, then we would expect that models trained mostly on the more easily findable DAAL would misrepresent the talk of users of AAL in their attempt to link language features and language varieties. In this work, we explore that probable misrepresentation and attempt to unmask how it arises, as well as its potential implications for AAL speakers whose language might be interpreted by these models.

3 Methods

This project uses a dataset consisting of 274 posts from a curated sample of texts across six different sources designed to simulate the kinds of African American Language training data which LLMs are exposed to. As described in [7][2], this selection of texts features regional and contextual variations of AAL: posts on the subreddit r/BlackPeopleTwitter (r/BPT); comments from the posts on r/BlackPeopleTwitter; tweets from the TwitterAAE corpus [2]; the Corpus of Regional African American Language (CORAAL, [21]); Columbia World Projects focus group data (CWP); and Hip-Hop Lyrics.

The two r/BPT datasets–posts and comments–consist of posts containing screenshots of materials created by Black people across social media (primarily Twitter/X)[3] and comment threads from those posts. The comments from the r/BPT dataset come from Reddit users who are members of the BPT Country Club,[4] which is a group of users on r/BlackPeopleTwitter who are verified by community mods as Black or POC (see [16]). The CORAAL datataset draws from transcripts of sociolinguistic interviews featuring regional AAL varieties from the Contemporary Corpus of Regional African American Language [21]. The Columbia World Projects (CWP) data consists of transcripts from 10 focus groups from a separately-funded sub-project of this team's overall work, wherein Harlem community members were invited to discuss loss and grief among the African American community in the wake of COVID-19. The Hip-Hop lyrics dataset contains lyrics from songs featured on

[2] Considering the potential misuses of our data (e.g. social media analysis for policing) as well as the potential for social good (e.g., improved representation of AAL in language technologies), we will make our data and annotations available to those that have signed an MOU detailing their research uses of the data and an ethics statement. See Deas et al. [7] for further discussion.

[3] The full description of the r/BlackPeopleTwitter subreddit reads "screenshots of Black people being hilarious on social media, it doesn't need to just be twitter but obviously that is best" (https://www.reddit.com/r/BlackPeopleTwitter/).

[4] As of this writing (June 2024), the post explaining the Country Club requirements has been deleted from r/BPT and the link to it in the community rules returns the link to a discussion of the post.

Billboard's 2022 Top Hip-Hop Artists charts as well as 27 songs, three each from nine Black artists in Morgan (2001). The subset of tweets from TwitterAAE were automatically detected using a demographic-aligned ensemble model that was trained to associate textual features with demographic data provided by the tweets' geo-tags. It's worth noting then that the TwitterAAE corpus privileges tweets from areas with large populations of African Americans, and is therefore unlikely to include tweets from users living in predominantly white areas. As such, the tweets are likely to contain features resembling AAL, but are not guaranteed to be created by African American users.

The datasets collected represent a variety of African American Language use. Some, (CORAAL, CWP) represent more naturalistic talk from African American speakers, and others (TwitterAAE, BlackPeopleTwitter) contain sources which are more likely to contain the kinds of overtly performative uses of AAL that are typical of DAAL and online performativity more generally. For the purposes of this work, we have separated the datasets into the performative datasets (r/BPT posts, Hip-Hop, TwitterAAE) and the non-performative (CORAAL, CWP, r/BPT comments.) It should be noted that while r/BPT comments, as coming from the users of r/BPT should theoretically be slightly less performative in nature than the tweets about which the posts are written, these users are nevertheless in an online environment and replying in-kind to posts which have been deemed by other users as particularly noteworthy, in other words, successful performances. Thus r/BPT comments might be considered somewhat a middle ground between overt performance and naturalistic speech. Finally, while most of the datasets consist of data collected independently of its features, the CORAAL data was collected via a regular expression evaluation of AAL features, which leads to a higher incidence of AAL than is common in naturalistic data; this has implications for our statistical analysis and comparisons as will be discussed below.

The texts from each dataset were separated into individual phrases ($N=274$). Next each phrase was coded by identifying the specific AAL features (or lack thereof) present in the text. For example, the phrase "Like tweetin around this time, aint nobody really up" (TwitterAAE) would be coded as containing a written fronted alveolar (ING) in "tweetin" (cf. tweeting) and the presence of *ain't i* (AI). Some phrases included multiple AAL features, so the data was cleaned by duplicating those entries such that each feature would be counted on its own, allowing us to get a clear picture of the frequency of individual features. In total 31 morphosyntactic and typewritten phonological features were coded.

Following [12, 13], our team looked for morphological, syntactic, and (typewritten) phonological features, rhetorical features (e.g. he ain't shit), and camouflaged language (e.g. *he came in here talking about X*, see [35, 36]. Interrater reliability was established at 92% between two coders. This comprehensive coding enabled us to look in aggregate at the overall rates of AAL use across our data, as well as to zero in on some specific features within that data. Below, we explore first the overall effect of the nature of the different datasets, and then explore some implications for understanding suggested by the patterning of individual features.

Table 1 Comparison of the most common feature, number of phrases with AAL, total phrases, rate of AAL use, and total number of individual features across each dataset

	Top feature	Number of AAL instances	Total phrases	Rate of AAL use	Total number of unique features
TwitterAAE Tweets	Other 16.66% ($n=17$)	102	104	98.08%	22
BPT Comments	Other 16.22% ($n=6$)	37	63	58.73%	16
BPT Reddit	Ain't 25.76% ($n=17$)	66	81	81.48%	18
CORAAL	Gon 16.66% ($n=8$)	48	59	81.36%	16
CWP	Ain't 18.75% ($n=6$)	32	51	62.75%	15
Hip-Hop	Ain't 24.69% ($n=20$)	81	83	97.59%	15

4 Results: Overall performativity

This section focuses on the breakdown of each dataset and the distribution of features within those datasets, beginning with a look at the presence or absence of AAL in each dataset. Table 1 shows the top feature, the total number of phrases, and the total number of unique features (i.e. codes) in each dataset (Fig. 1).

Comparing the performative and non-performative datasets directly, we can see differences between AAL use across the spectrum of posts produced. The TwitterAAE and Hip-Hop lyrics datasets have the highest rates of AAL features at 98.0% and 97.6%, respectively. This means that less than 3% of all phrases coded from those datasets lacked AAL features. Notably, data gathered from Twitter (i.e. TwitterAAE and BPT Reddit posts) also feature the widest variety of AAL features, with the TwitterAAE dataset showing the highest number of AAL features both as a raw count and percent of total phrases Given the overt performativity of Hip-Hop lyrics and the ways in which AAL is leveraged on

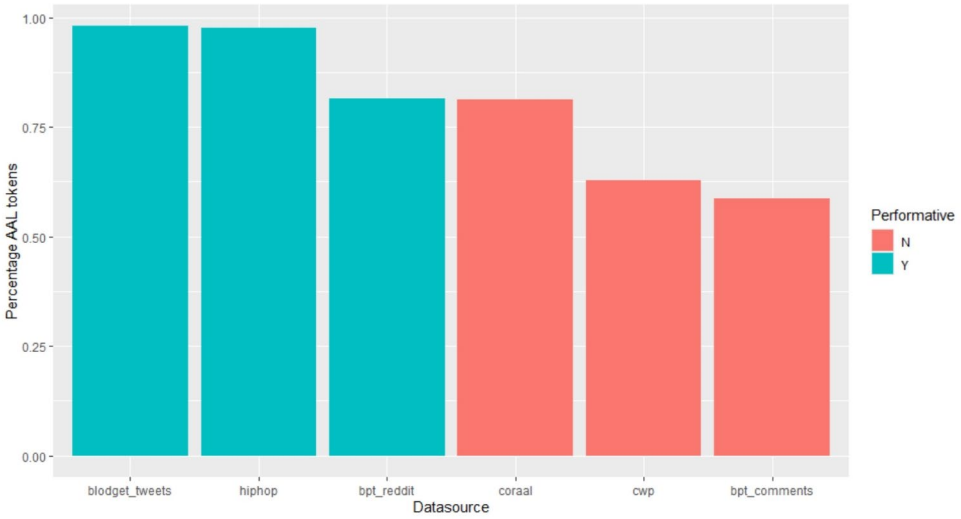


Fig. 1 Comparison of the performativity of datasets and the percent of AAL phrases in each dataset

Table 2 Examples of features of AAL present in datasets

Feature	Example
ain't (AI)	If that **ain't** the truth (bpt_reddit148930)
orthographic capture of phonological features of AAL (PP)	**Ova** my shoulda (blodget_tweets482050)
other nonspecific features of AAL (OT)	If it ain't i Message **it ain't shit** (bpt_reddit188909)
unexpressed / zero copula (ZC)	Then she **Ø** drinking wine too (blodget_tweets497159)
negative concord (NE)	It ain't **no** fun, if the homies **can't** have **none** (bpt_reddit7186)

Twitter for potential virality, this suggests that environments where performativity is a goal will see not only more AAL features employed overall, but also a greater variety, signaling the importance of AAL repertoire to the performative linguistic style of these communities.

An independent t-test of the means of the AAL rates between the two groups is significant at $p = 0.01$. We choose here to exclude the CORAAL data, which, as discussed in the methods section, was collected to include only phrases which contained AAL features.

5 Results: Features

We further zeroed in on specific features of AAL present in the data to see if there were differences in the representation of individual features across the various datasets, as related work by our team has shown that some specific morpho-syntactic features of AAL pose a greater challenge for LLM interpretability than do others (Grieser et al. Under Review). We selected the five most commonly-occurring features across the data, which also happened to be all the features which had a total N of at least 20 instances across all six datasets. These features, with representative examples from the data, are presented in Table 2. A sixth feature, ING, or dropping of the orthographic <g> at the end of words ending in *-ing* was tied with negative concord for the fifth most commonly occurring feature, however, this feature was not coded in the CORAAL and CWP data, which are spoken; a future comparison of spoken data and orthographic data would provide additional insight into the naturalness of online AAL (Table 3, Fig. 2).

While some domains feature a higher prevalence of AAL, some general trends can be noted from this by-feature data. For instance, Hip-Hop lyrics show a higher presence of features associated with the sound system, likely due to the fact that this domain is performed audibly. On the other hand, zero copula, one of the principal features that distinguishes AAL from mainstream white varieties of English, is relatively uniformly represented across the different datasets. We can also see differences across the datasets–the CORAAL and CWP datasets, which consist of naturalistic spoken data, have the most even distribution of different features.

Table 3 Top five most common features present in all datasets

	TwitterAAE	BPT Comments	BPT Reddit	CORAAL	CWP	Hip-Hop	Total
NE	3 (2.94%)	2 (5.41%)	5 (7.58%)	3 (6.25%)	1 (3.13%)	7 (8.64%)	21 (5.73%)
ZC	8 (7.84%)	1 (2.70%)	5 (7.58%)	4 (8.33%)	5 (15.63%)	4 (4.94%)	27 (7.37%)
OT	17 (16.67%)	6 (16.22%)	9 (13.64%)	4 (8.33%)	2 (6.25%)	7 (8.64%)	45 (12.29%)
PP	16 (15.69%)	5 (13.51%)	7 (10.61%)	5 (10.42%)	2 (6.25%)	11 (13.58%)	46 (12.58%)
AI	12 (11.76%)	5 (13.51%)	17 (25.76%)	6 (12.50%)	6 (18.75%)	20 (24.69%)	66 (18.0%)

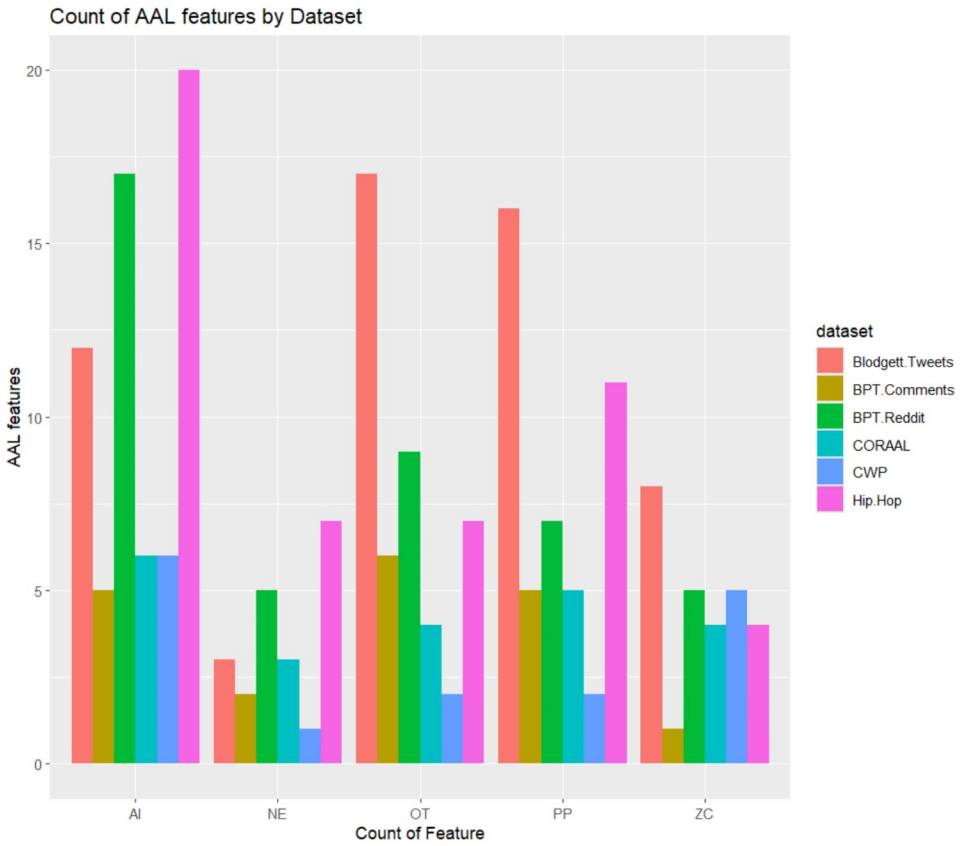


Fig. 2 Count of top five most common AAL features by dataset

It is worth noting that features that are part of AAL may not necessarily be unique to AAL, and there can be overlap between features present in AAL and features present in other less-privileged varieties of English. For example "ain't" shows up in AAL but it is also present in varieties of southern English (notwithstanding the overlap of AAL and southern English and their influence(s) on each other). This cross-dialectal use of *ain't* may explain its significant presence across all six datasets; the use of it can cross-cut different audiences and signal the same performative aspect to all groups without specifically marking an African American group.

In conclusion, the aggregate and by-feature results demonstrate that while specific features may vary, there are overall significant differences in both the prevalence of features and the features used across the non-performative and performative datasets. The non-performative datasets show a lower rate of AAL feature use and a more even distribution of specific features across the data, while the performative datasets not only demonstrate a higher rate, that rate is composed of a more compact set of features.

The data presented here is compiled to mimic the kinds of training data to which many LLMs are exposed. Because of the proprietary nature by which companies such as Meta, Alphabet, Microsoft, and Open.AI control the use of their training data, we cannot conclude definitively that the demonstrated biases in the models produced by these organizations are certainly or entirely attributable to a misrepresentation of AAL in the training data. Nevertheless, our analysis suggests that when LLMs use easily-obtained scraped training data, the performative nature of digital AAL means that LLMs are likely to train in a representation of African American Language users which over-represents the use of AAL features as a raw rate, while simultaneously under-representing the diversity of the kinds of language features used by real speakers of the variety.

6 Discussion

Given the marked rise of and investment in AI, it is vital to address the shortcomings of emerging technologies, such as LLMs. This discourse is especially important given the history and context of racially biased LLMs and social media platforms being used as tools of surveillance, resulting in repercussions both online and in the physical world. Previous literature has shown that AAL speakers are viewed as less credible and less educated more generally [22]. LLMs, designed by humans who often unconsciously hold these biases and trained on data which reflects these biases, are

more likely to flag innocuous AAL content as toxic or violent. When content is posted to perform a version of one's identity, voyeurs on social media, who are outside the intended audience (such as law enforcement officials), may rely on LLMs to interpret said content and can ultimately misinterpret meanings and intentionality. Stigmatized language varieties are already susceptible to misinterpretation; for example, previous literature shows that jurors are biased against AAL speaking defendants, and can even misinterpret facts of a case that were expressed in AAL [22]. The implication that LLMs can view certain content–such as Hip-Hop lyrics or use of slang–as particularly incendiary language can add to an already precarious world where Black youth are heavily policed online, and can face potential incarceration, suspension, and termination from jobs due to online performance [29]. This online over policing is an extension of the long history of AAL being misinterpreted by non-Black people with negative results for Black people.

A secondary implication of the misinterpretation of performance and camouflage online is the importance of comparing posters and voyeurs…and contend with the potential for virality. #BlackTwitter has been a hotbed for discussion of Black cultural moments, humor as a response to current events, and political activism and racial justice [25]. Additionally, since tweets from #BlackTwitter don't always *stay* on #BlackTwitter, the white gaze is just as omnipresent online as it is offline [5]. In the context of digital performativity, the identity of individuals posting is salient to understanding the meaning of their posts. Succinctly, it matters whether a Black person is using features of AAL online, or if a white person is. The *meaning* of certain phrases in AAL are entirely contingent upon a rich cultural context and understanding that exists within Black communities–and when co-opted by white communities–those same phrases may lack contextual nuance. #BlackTwitter users may perform in ways that are vastly different than white Twitter users, and if LLMs are not aware of how to pick up on what we term as unmarked performativity, or performance that is only understood as subtle nuances by humans, there is a risk of further digital cultural appropriation, or what Joshua Lumpkin Green coined as digital Blackface [11]. The same issues would be expected with respect to other minoritized language varieties, such as Multicultural London English (MLE), and an exploration of how LLMs interpret minoritized language in other countries is a direction for future research.

It has also been studied that when AAL is co-opted by white communities, the meaning of certain phrases often changes [19].Thus, a similar pattern may occur, where LLMs can favor the meaning implicit in white versions of language, versus what it means in AAL. Moveover, it is necessary to consider the ethics of taking data from communities like #BlackTwitter to use to train LLMs at all, let alone to develop technology that may serve to further marginalize those same groups [23]. Performing identity on social media has multiple different functions: although speakers of AAL, primarily Black speakers in the U.S., are often forced to code switch and effectively perform for a white audience in predominantly white spaces, digital communities such as #BlackTwitter exist as a counterpublic outside of white voyeurism [10]. Posts using AAL on #BlackTwittter utilize a linguistic creativity that resists standard linguistic functions–and reflects the argument that #BlackTwitter is not only a space for Black joy, but also a space to question the white norm. By non-consensually taking data from these spaces in order to train language models, LLMs become the white gaze that may appropriate and misinterpret content created by Black authors.

7 Conclusion and implications

The probabilistic nature by which LLMs interpret training data means they lack the ability to understand what is embellished for the sake of depicting a 'spatial self' that reflects cultural knowledge and awareness of the intended audience [34, p. 3]. Instead, they are likely to interpret the language patterns in their input data as being emblematic of a particular kind of language user. As a result, LLMs are highly likely to misinterpret content and therefore flag innocuous content as toxic or violent when it comes from a marginalized langauge variety like AAL. This paper aims to discuss a large implication of LLMs and misinterpretation: there are a myriad of harmful outcomes of AAL being misunderstood by voyeurs of social media posts who are external to the posters' intended audience, particularly law enforcement and adjacent institutions involved in the hyper-surveillance of Black people online.

Lane [24] discusses Twitter as a "digital street," where threats and other gang violence can play out digitally and exude into offline altercations (p. 163). Performativity is inherent on the digital street, the same way it is inherent on the real one–particularly when people need to posture as strong, confident, or untouchable for survival. For instance, Black youth living in communities with high levels of gun violence may post imagery with guns as a way to posture and display masculinity, bravery, and aggression, but may not actually have intent to utilize guns or violence [29]. Social networking sites are increasingly policed and surveilled, and without a firm foundation in what tweets may lead to violent behavior and what may be social posturing, police presence and hyper-surveillance can cause incredible harm and lead to incarceration [29]. Given the bias-inducing data skew which we have explored in this work, the

increasing use of LLMs and other artificial intelligence programs in understanding online space thus poses a significant and disproportionate threat to people of color, and this threat must be carefully understood by those who would use–and especially those who would create–these tools.

Finally, this work raises a provocative question: given that the performative nature of online language results in a misrepresentation of African American Language, should LLMs be better trained to interpret this performativity? Although it is important to understand and attempt to mitigate potential causes of harm, current practices for training LLMs on language that likely misrepresents the nature of language use by African Americans means that it will be difficult, if not impossible, to develop models that have the cultural context necessary to properly interpret that language. LLMs currently exist in a socio-political and historical vacuum, while language does not. Without the ability to understand linguistic output as being rooted in systemic oppression, cultural expression, and resistance within the fabric of race in American life, current LLMs are not equipped with the tools necessary that humans have to fully interpret AAL and therefore to unproblematically understand its users.

Although LLMs are developing and morphing at an incredibly fast rate, at present much of their training is based on an assumption that the language upon which they are trained accurately represents the language use of the communities the language represents, an assumption which this work shows is likely faulty. At the same time, models which entirely exclude the language of minoritized speakers in their training further marginalize those speakers when tasked to interpret their language. Given that language use reflects the intention of its speakers and the norms of their speech communities in ways that even other humans often find difficult to interpret, is it possible to train LLMs to understand the ways in which their own training can reflect human biases and intention? Moreover, what would it take to make AI technology more capable of taking culture and social context into account? We argue that the modeling of language to accurately and unproblematically interpret the language of minoritized speakers will continue to pose a major challenge to the AI field, and current approaches require substantial changes to enable even a fraction of the understanding that humans achieve.

Acknowledgements This work was supported in part by grant IIS-2106666 from the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. This material is also based upon work supported by Columbia World Projects, under a grant titled, *Tools and Connections: Strengthening Support for COVID-19 Bereavement in Black Communities in Harlem.*

Declarations

Conflict of interest The authors whose names are listed immediately below certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Bell, A.: Language style as audience design. Lang. Soc. **13**(2), 145–204 (1984)
2. Blodgett, S.L., O'Connor, B.: Racial disparity in natural language processing: a case study of social media African American English. Cornell University (2017). arXiv https://doi.org/10.48550/arxiv.1707.00061
3. Blodgett, S.L., Green, L., O'Connor, B.: Demographic dialectal variation in social media: a case study of African American English. Cornell University. (2016). arXiv https://doi.org/10.48550/arxiv.1608.08868
4. Cunningham, J.: Features of digital African American language in a social network site. Writ. Communication. **31**(4), 404–433 (2014). https://doi.org/10.1177/0741088314547188
5. Daniels, J.: Race and racism in internet studies: a review and critique. new. Media Soc. **15**(5), 695–719 (2013)
6. Davidson, T., Bhattacharya, D., Weber, I.: Racial bias in hate speech and abusive language detection datasets. Proc. Third Workshop Abusive Lang. Online. (2019). https://doi.org/10.18653/v1/w19-3504
7. Deas, N., Grieser, J., Kleiner, S., Patton, D., Turcan, E., McKeown, K.: Evaluation of African American language bias in natural language generation (2023). arXiv https://arxiv.org/abs/2305.14291
8. Florini, S.: Tweets, tweeps, and signifyin': communication and cultural performance on Black Twitter. Telev. New. Media. **15**(3), 223–237 (2014). https://doi.org/10.1177/1527476413480247
9. Goffman, E.: The presentation of self in everyday life. 1959. Garden City, 259 (2002)
10. Graham, R., Smith, S.: The content of our #Characters: Black Twitter as counterpublic. Sociol. Race Ethn. **2**(4), 433–449 (2016). https://doi.org/10.1177/2332649216639067
11. Green, J.L.: Digital blackface: the repackaging of the Black masculine image. Master's Thesis, Miami University (2006)
12. Grieser, J.A.: Investigating topic-based style shifting in the classic sociolinguistic interview. Am. Speech: Q. Linguistic Usage. **94**(1), 54–71 (2019)
13. Grieser, J.A.: The Black Side of the River: Race, Language, and Belonging in Washington. Georgetown University Press, Washington (2022)

14. Grieser, J.A., Shepard, J., Deas, N.: (Under Review) Exploring the deficiencies of large language models in use of AAL: an interdisciplinary study. Under review at Linguistic Vanguard
15. Hamel, E., Kani, N.: Factors that influence automatic recognition of African American vernacular english in machine-learning models. IEEE/ACM Trans. Audio Speech Lang. Process. **32**, 509–516 (2024). https://doi.org/10.1109/TASLP.2023.3331139
16. Harmon, A.: Discussing Blackness on Reddit? Photograph your Forearm First. New York Times (2019). (nytimes.com)
17. https://www.kaporcenter.org/wp-content/uploads/2023/03/KC22001_black-tech-report.final_.pdf
18. Ilbury, C.: Sassy Queens: stylistic orthographic variation in Twitter and the enregisterment of AAVE. J. Sociolinguistics (Print). **24**(2), 245–264 (2019). https://doi.org/10.1111/josl.12366
19. Jones, T.: Toward a description of African American vernacular english dialect regions using Black Twitter. Am. Speech. **90**(4), 403–440 (2015). https://doi.org/10.1215/00031283-3442117
20. Kapor Center: State of Tech Diversity: The Black Tech Ecosystem (2023)
21. Kendall, T., Farrington, C.: The Corpus of Regional African American Language. Version 2023.06. Eugene, Ore. The Online Resources for African American Language Project (2023)
22. King, S., Rickford, J.R.: Language on trial. Dædalus. **152**(3), 178–193 (2023)
23. Klassen, S., Fiesler, C.: This isn't your data, friend': Black Twitter as a case study on research ethics for public data. Social Media + Soc. **8**(4), 205630512211443 (2022). https://doi.org/10.1177/20563051221144317
24. Lane, J.: The Digital Street. Oxford University Press (2018)
25. Lockett, A.: What is Black Twitter? A rhetorical criticism of race, dis/information, and social media. Race, Rhetoric, and Research Methods, 165–213 (2021)
26. McCulloch, G.: Because Internet. Riverhead Books, New York (2019)
27. Mitchell-Kernan, C.: Signifying, loud-talking, and marking. In Dagel Caponi, Gena, ed. Signfyin(g), Sanctifyin', & Slam Dunking: A Reader in African American Expressive Culture. University of Massachusetts Press, Amherst (1999)
28. Papacharissi, Z.: Without you, I'm nothing: performances of the self on Twitter. Int. J. Communication. **6**, 18 (2012)
29. Patton, D.U., Eschmann, R.D., Elsaesser, C., Bocanegra, E.: Sticks, stones and Facebook accounts: what violence outreach workers know about social media and urban-based gang violence in Chicago. Comput. Hum. Behav. **65**, 591–600 (2016). https://doi.org/10.1016/j.chb.2016.05.052
30. Patton, D.U., Frey, W.R., Gaskell, M.: Guns on social media: complex interpretations of gun images posted by Chicago Youth. Palgrave Commun. **5**, 119 (2019). https://doi.org/10.1057/s41599-019-0330-x
31. Patton, D., Darkwa, E., Anguiano, K.: Recommendations to end virtual stop and frisk policing on social media. Tech Policy Press. Accessed (2021). https://www.techpolicy.press/recommendations-to-end-virtual-stop-and-frisk-policing-on-social-media/
32. Pavalanathan, U., Fitzpatrick, J., Kiesling, S.F., Eisenstein, J.: A multidimensional lexicon for interpersonal stancetaking. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, vol. 1: Long Papers, pp. 884–895 (2017)
33. Rightler-McDaniels, J.L., Hendrickson, E.M.: Hoes and hashtags: constructions of gender and race in trending topics. Social Semiot. **24**(2), 175–190 (2013). https://doi.org/10.1080/10350330.2013.859355
34. Schwartz, R., Halegoua, G.R.: The spatial self: location-based identity performance on social media. New. Media Soc. **17**(10), 1643–1660 (2015). https://doi.org/10.1177/1461444814531364
35. Spears, A.K.: The Black english semi-auxiliary come. Language. **58**(4), 850–872 (1982). https://doi.org/10.2307/413960
36. Weldon, T.L.: Middle-class African American English. Cambridge University Press (2021)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

AI and Ethics
https://doi.org/10.1007/s43681-024-00626-z

ORIGINAL RESEARCH

From human-system interaction to human-system co-action and back: ethical assessment of generative AI and mutual theory of mind

Florian Richter[1]

Received: 27 June 2024 / Accepted: 18 November 2024
© The Author(s) 2024

Abstract
Human-machine ethics has emerged as a rapidly growing research field in recent years. However, it seems that Generative Artificial Intelligence (AI) leads to a paradigm shift from human-machine interaction to co-action. The ethical assessment of such relationships is still in the making and needs further scrutiny. First, studies about the influence of technology in human-system interactions and manipulation are reviewed. Second, the "mutual theory of mind" approach is critically examined to identify its shortcomings. Third, creating user models is reconstructed to demonstrate the strategies of systems. Finally, use cases are discussed and assessed to outline ethical implications.

Keywords Generative AI · Mutual theory of mind · Human-machine ethics

1 Introduction

Technology ethics has become a multifaceted research area ranging from normative to empirical investigations. Specifically, in human-machine ethics, it is paramount to investigate technology's mediating role and ethical implications. The conceptual level has been investigated, for instance, in postphenomenology [1–3]. Furthermore, in recent years, several studies have been conducted to shed light on the influence of technology on human behavior in human-machine ethics, from the influence of chatbots [4] to the ethics of autonomous vehicles [5, 6], hybrid traffic [7], and the ethical implications of care [8]. However, immersion in virtual reality and real-time feedback from systems to subjects via Generative Artificial Intelligence (AI) is still in the making. It is discussed under the concept of co-action. The ethics of human-system interaction, which AI technologies have already disrupted, faces new challenges in developing ethical assessment tools for human-system co-action.[1]

In human-system interactions, the developers and users try to align their expectations about how the artifact/system should function and how it should be designed. This is done via studies in UX Design, acceptance research, etc. To a certain degree, the *practice* is more or less transparent. In human-system co-action, the user *and* the (semi-) autonomous system produce an effect or output. Examples are text or picture generation via prompts or brain-computer chips. However, it is opaque to the user how the system functions to produce a specific effect and who is responsible in the end. How is it possible to reflect on the role of the system in creating the output? How can the user be involved in the developmental process? And how can such systems be assessed ethically. It is hypothesized that if the system functioning is made explicit to yield a more transparent practice, like in human-system interaction, and the user is included, then an ethical assessment of co-actions and the use of Generative AI is possible. Therefore, we must move from human-system interaction to human-system co-action *and back* [9, pp. 140–144]!

Aligning expectations of systems such as conversational agents (CA) and humans has been discussed under a "mutual theory of mind" [10, 11]. It should allow "smoother human-CA conversations" by adapting the mutual theories or models of the mind [10]. However, it is unclear if such models of the mind reflect either the system's functioning or the human agent well. Instead, the system ascribes roles, creates relevance rankings, establishes routines for the user,

[1] I am grateful for comments from Matthias Uhl and many discussions with him about the topic and an earlier draft of the paper.

✉ Florian Richter
florian.richter@thi.de

[1] Technische Hochschule Ingolstadt, Esplanade 10, 85049 Ingolstadt, Germany

Published online: 04 December 2024

and based on such an analysis, the system offers recommendations, explanations, or options for the user. By having an actual theory of mind, the user can also, to a certain extent, exploit the system's strategic functioning to reach a specific goal or objective. This could be extended by assuming that the system or the virtual conversational agent has adaptational strategies, too, to react to these strategies. The strategic "exploitation" could be captured via game theoretic models. Nevertheless, it presupposes that the user and the system "understand" each other, i.e., can interact on a strategic level. However, in the case of co-action, where an effect is produced by the user and a (semi-)autonomous system [9, p. 140], like in the generation of pictures or texts via prompts or in the use of brain-computer interfaces, difficulties arise to align the expectations of the user with the system strategies due to the opaqueness of the action environment and context for the user. This has various ethical implications regarding how trust can be developed and how the system strategies influence the users without manipulating them.

The paper makes four contributions. First, studies about the influence of technology in human-system interactions are reviewed to assess and summarize the ethical implications of the mediating role of technology. Second, the approach of a "mutual theory of mind" is discussed. Its shortcomings are identified in opening the search space for more adequate tools to assess human-system co-action. Third, creating user models is analyzed to demonstrate the strategies of systems. However, and that is the final aspect of the paper, such strategies can, on the one hand, be exploited by users but, on the other hand, be used to control or manipulate the user. Here, use cases are discussed and assessed to outline ethical implications.

2 The mediating role of technology

2.1 Philosophical background

Technology is sometimes defined as a collection or system of connected artifacts [12, p. 42, 13, p. 2]. This usually entails an instrumental reduction of artifacts serving as tools to achieve goals set by humans [14]. To address these reductionist views, various approaches have been proposed in recent decades, such as ideas that stem from Wittgenstein's philosophy of language [15], postphenomenological approaches (e.g., Don Ihde and Peter-Paul Verbeek), dialectical approaches [16], or lines of thought that emerged from Technology Assessment [17]. All approaches have in common that technology mediates our access to the world and us in a certain way. Technologies are not "opposed to humans, or mere extensions of us, they need to be seen as media for our connections with the world" [1, p. 29]. (This presumably explains also the intricate relation of the philosophy of technology with anthropological statements like, e.g., that we are "deficient beings" (Arnold Gehlen).)

Each form of mediation entails a certain ethical viewpoint on technology. From an instrumentalist perspective, artifacts appear just as means and are value-neutral or -free [14]. From a postphenomenological perspective, technology seems to be shaping our access to the world. Consequently, technology is imbued with values [3]. Despite the fierce defense of the value-neutrality thesis of technology [14], it can be doubted that it can serve as a basis for technology ethics and adequately grasp its challenges. Technological systems can be much more complex than tools like guns or hammers. Some systems are highly automated and act partly autonomously (like self-driving cars). Hence, they could be understood as displaying, to a certain degree, moral agency (compare the four levels of moral agency by James H. Moore [18]). Additionally, some systems have a high depth of engagement into the human (e.g., chips implanted into the brain). For instance, self-driving cars operate within a complex system and interact with humans (pedestrians, human drivers).

2.2 Empirical studies in human-machine ethics

Therefore, based on the theoretical assumption that technology is value-laden, shapes our practices, and forms who we are as humans, several empirical studies have been conducted to shed light on the influence of technology on human behavior in human-machine ethics. Some approaches stem from behavioral economics or psychology [4–8]. Although empirical studies about the influence of moral advisors are still scarce [19], some studies show that users follow moral advice from systems or bots even if it is clear to them that they are interacting with an AI-based system [4, 20–22] Furthermore, it seems that people sometimes overtrust robots (for instance, in emergency situations) [23]. Hence, it seems evident that technological systems can influence the moral behavior of humans.

Under the assumption that a clear moral framework could be provided, it would, therefore, seem that a certain kind of nudging via technological systems could also be in the interest of its users [24]. The system would influence the users positively, helping them to make morally correct choices or to live in a morally right way. Although Thaler and Sunstein proclaim that nudging people is just a "paternalism of means" [25, p. 7 and 312], it is, in fact, also a paternalism of ends because some kind of common sense moral framework is supposed that neither is questioned nor does it include stakeholders to assess which ends they consider acceptable.

Klincewicz explicitly considers moral AI machines as "means" to influence behavior and to somehow remove immoral behavior. He writes, "it does give a promising alternative to pharmaceuticals and neural intervention for individualized moral enhancement." Furthermore, such a moral enhancement would not be paternalistic, according to him, because it persuades reason-responsive beings and is thus not coercive. Nevertheless, the problematic issue is that Klincewicz already knows what the morally correct behavior is, at least in some specific cases that he discusses. In his example, a racist would and should be persuaded by a moral machine to report an abuse of a Chinese person by a police officer [26]. Yet, this is a moral problem and not an ethical issue where different ethical stances, values, and principles must be weighed against each other. To extend a moral machine that solely functions as a persuasion tool to these ethical cases would then not only be a paternalism of means and nudge people to do the morally right thing, but it would also be a paternalism of ends.

The example from Klincewicz above does not need specific technology ethics. General moral considerations would be sufficient [27, p. 16]. However, it is crucial to investigate how AI-based systems influence their users and co-agents, i.e., how they shape the space of possibilities to (co-)act. In this process, two problematic issues arise:

1. It is essential to distinguish between mere influence that can be considered ethically acceptable (like most advertisements) and manipulation (trickery like grandparent scam). Here, clear normative criteria need to be established.
2. Biases and discrimination are essential parts of the debate in the ethics of algorithms [28]. However, biases cannot be eliminated completely to reach an unbiased original position.– Technology (but also morality) would be value-neutral.– Any relevance ranking already discriminates in the sense of distinguishing between something; otherwise, we could only use meaningless tautologies, such as "Manipulation is manipulation." or "A value is a value."

2.3 Technology between influence and manipulation

While the former issue will be examined here shortly, the latter will be discussed in section four. Although "it has been suggested that manipulation sits on a continuum of influence between rational persuasion and coercion" [29, p. 4], it is rather difficult to determine clearly what manipulation is. Some claim that to "covertly influence" someone is manipulation, i.e., "to intentionally alter their decision-making process without their conscious awareness" [30]. However, manipulation does not need to be hidden; for instance, a real estate agent might offer freshly baked cookies in a house, while the potential buyers might also be aware of the fact that they are manipulated [29]. Another attempt is to define manipulation as "bypassing the chooser's deliberative capacities" [31, p. 86]. Nevertheless, one might also be manipulated by responding to reasons. For instance, one might have a strong sense of fairness, and a manipulator might appeal to it to get a better share. Furthermore, manipulation is characterized as trickery that presupposes an intention on the side of the manipulator [32]. However, it can also be the case that the person is a pathologic manipulator and manipulates without being consciously aware of it. Additionally, despite not having intentions, technological systems could be characterized as manipulative. Even though such systems might be used solely as tools by a human agent to manipulate someone, some systems can also be manipulative without being intentionally designed to do so [29, p. 8]. Manipulation does not need to be done intentionally. As long as it is goal-oriented (where the goals are either set by humans or systems), it could be manipulative [33]; otherwise, "influence that is purely accidental" would count as manipulation [29, p. 9].

Manipulation is non-intentional [29] and is thus different from concepts like nudges, dark patterns, and persuasive technology.[2] In the latter concepts, some kind of intentional and deliberative decision on the side of the designer or developer is made to influence the user, whether in a good or bad way. Nudges are, for instance, based on a choice architecture that is deliberately constructed to influence the user to achieve a goal that supposedly– according to the designers or developers– is good for her (for an overview of the topic, see [34, 35]). Dark patterns, on the other hand, "are deliberately constructed designs with the intent of inducing users to take actions they might not willingly choose if they were fully informed or aware of alternatives" [36, p. 207, 37]. Representants of libertarian paternalism consider nudges ethically acceptable [24], while dark patterns are considered ethically unacceptable. Persuasion could be considered acceptable if it is based on (good) reasons, while it can be considered unacceptable in the sense of coax or cajole. Of course, such a distinction is very rough, and there is also a "tension between persuasive UX practices and manipulative designs" [38]. Nevertheless, it seems that normative criteria are established to a certain degree in these cases within the field of human-computer interaction. However, manipulation is rather difficult to determine ethically.

[2] I am thankful to an anonymous reviewer for pointing out that these concepts play a crucial role in human-computer interaction. Nevertheless, the main background of the paper is philosophical and ethical and thus rooted more in the metaethical discussion.

Klenk adds what he calls an "indifference criterion": the manipulator is "indifferent to revealing reasons to their victims in their choice of the means of influence that they employ." Furthermore, it "can be interpreted non-intentionally by thinking about the function of a chosen means of influence," like in a movie recommender system [29, p. 8/9]. Nevertheless, this seems to presuppose some strategic planning on the side of the manipulator, which means it should be chosen to achieve a goal most efficiently and effectively. The problem is here that almost any kind of influence could count as manipulative. A driver assistant system that influences the driver on an operational level would also be manipulative as long as it does not reveal reasons why it chooses certain means to protect the driver. Usually, there is also no time for some deliberative process or exchange with the driver to give her reasons, which means should be chosen regarding the situation.

Yet, the systems were tested and developed accompanied by a deliberative process. The debate's primary focus is evaluating means-end schemata that individuals could employ. Ultimately, the ethical task would be to assess the manipulative side-effects of goal-oriented behavior and their risks. Also, Klenk already sees the problematic issue that systems can be manipulative and that questions of manipulation must be incorporated into the technological systems design process to prevent manipulation [29]. However, the system perspective must be more clearly elaborated by investigating the system strategies employed to model users and how users can be included in shaping the systems. This will also have the advantage that worries where influence "undermines the target's autonomy" because "first, it can lead them to act toward ends they haven't chosen, and second, it can lead them to act for reasons not authentically their own" [30, p. 9], could be disregarded. The stakeholders can commit themselves to system strategies that serve them and agree to them in specific contexts, i.e., they consider them (ethically) acceptable.

The advantage is that criteria for the acceptability of influence can be established. At the same time, manipulation can be used for specific moral cases, such as manipulating someone to act against her values. The former is an ethical issue from a system perspective, whereas the latter is a moral one and within personal ethics. The next step will be to examine an attempt to understand system strategies ("mutual theory of mind"). However, it will be shown that such an approach is insufficient as insofar as the systems are understood as intentional agents.

3 Aligning expectations

3.1 Mutual theory of mind in human-computer interaction

The ontological question of whether machines can possess consciousness is rather speculative. Nevertheless, it might be that users perceive machines as conscious. A study demonstrated "that people already ascribe a degree of consciousness to existing technologies" [39] Furthermore, in a Theory of Mind task, where participants had to describe something to a listener, here a virtual agent, it has been found that the participants adapted "their response by 30% when conducting the Theory of Mind task with a virtual agent compared to doing the task alone" [40, p. 6]. Nevertheless, it is rather one-sided ascription of consciousness than a mutual theory of mind.

Aligning expectations of systems like CA and humans has been discussed by Wang et al. under a "mutual theory of mind." It should allow "smoother human-CA conversations" by adapting the mutual theories or models of the mind [10]. However, it is not so clear what the goal of Wang et al. is. It seems like they want to confront the high expectations of users so that they adjust (and maybe lower) their expectations of the system functioning [9] because, as studies suggest, users have high expectations of CAs [41, 42] and this can lead to disappointments or to stop using such systems. It would thus be essential to design systems that use affordance cues to clarify the system's capabilities [43]. According to the study of Wang et al., a personalized and better-adapted functioning (in their study: better answers) seems to correlate with an assessment of the system as presenting a more human-like and intelligent picture of the system [10]. Of course, better answers can come along with more benevolent attitudes. Furthermore, they infer that someone with a more verbose language might "confuse" the CA with complex inputs, so it might fail to deliver "supportive or efficacious responses, leading to undesirable CA perception" [10]. However, someone with a broader vocabulary would probably not consider a CA an adequate interlocutor and thus have an undesirable perception of the CA. Assuming that either the CA has a theory of mind for the human user or the other way around seems inadequate because it is mainly based on stereotypes (heuristics) that each one creates.

Interestingly, such heuristics can also be biased, as described by Eicher et al., who start with the assumption that the student "has a theory of mind for how the computer operates." However, common "misconceptions," for instance, in variable assignments in programming, can be seen as systematic departures or biases about the functioning of the computer. They identify a set of common misconceptions to develop a "tool […] to give the computer the

ability to develop a theory of mind about the student and to use that in collaboration with the student to correct misconceptions" [11]. It might be doubted whether something like this can be interpreted as a (mutual) theory of mind where the tool develops a model for the student, and the student has a (sometimes biased) model of the computer. Instead, it is the assignment of a stereotype to the user/student and offering a recommendation or helpful explanation. By having an actual theory of mind, the student can also, to a certain extent, exploit the strategic functioning of the program to obtain instant feedback without a more extensive quarrel with her mistakes.

This could be extended by assuming that the system or the virtual CA has adaptational strategies, too. Hence, future studies in the field of ethics of Generative AI should also consider game theory to investigate how humans and systems employ strategies to get what they want, need, and seek depending on their motivations or programmed structures. This point also depends on the design of the chosen architectures.

Some literature about strategic behavior and a theory of mind for technological systems is located within the fields of human-robot or robot-robot interaction, where, for instance, one robot predicts the movement of another robot [44]. One study is based on a version of the prisoner's dilemma where humans interact with robots [45]. Furthermore, it has been investigated how a shared plan between a robot and a human can be executed without too many disturbances by the robot. Such interruptions by the robot could potentially lead to annoyance on the side of the human. For that purpose, the robot monitors the human agent and builds some kind of theory of mind [46]. However, such attempts to suppose a theory of mind remain within operational or strategic behavior. So, although the theory of mind covers the ability to "understand" shared plans, strategies, or predicting movements, it is far away from a theory of mind. Different topics are usually discussed in the philosophy of mind and are connected with the concept of mind, like, for instance, intentions, feelings, consciousness, critical judgment, etc. These abilities or characteristics are usually ascribed to human agents but not to technological systems. Scott et al. also state, "Theory of Mind is a specialist term." In their survey of 100 people about "conceptualizations of or attitudes towards machine minds," they found that "not one participant referred to Theory of Mind" [39, p. 5]. Furthermore, human agency is not limited to strategic or reward hacking behavior to find shortcuts. Human agents can commit themselves to a certain behavior or propositional content, understand their behavior's (moral) consequences, and give reasons for it. They are able to play a normative-deontic game [47].

Large language models are also used to assess human interpretation of robot behavior [48]. Deshpande and Magerko try to use "social cognition" and the "Observable Creative Sensemaking (OCSM) framework" to improve human-system interactions [49, p. 4]. They even use Generative AI, particularly image generation, as a "speculative case study". They state that the integration of such a framework "into Midjourney" would enable "AI to participate actively in the creative process. The model becomes more adept at interpreting creative prompts, engaging in a dynamic dialogue with users, and producing images that are not only contextually relevant but also aligned with the user's evolving creative journey" [49, p. 5]. Through "dialogue with users", "user feedback", and the generation of "multiple image options" a better understanding of human interactions with systems could be achieved [49, p. 5]. However, Deshpande and Magerko do not reflect the ethical dimension of all these issues and how, already in the developmental phase, human-system interaction could be made more transparent. Thus, feedback while using the system remains more within the co-action paradigm and might still be rather opaque for the user.

3.2 Aligning expectations in cognitive psychology

Ascribing systems a theory of mind (either to have a mind or to understand the mind of humans) is thus just a metaphorical way of speaking, and so is to speak of cooperation between humans and technological systems. How humans align and develop to have expectations has been investigated empirically in cognitive psychology. We can switch "perspectives as needed indefinitely," and we have an "ability" that is "involved in many aspects of shared intentionality": "recursive mindreading or recursive intention-reading"[3] [50, p. 96/97]. However, there is also a normative layer that must be reflected to fully understand expectations:

This "recursive intention reading transforms everything: turning helping and sharing into mutual expectations or even norms of cooperation." Through these mutual expectations, we bind ourselves to norms. That is why it is "turning practical reasoning into cooperative reasoning" [50, p. 106].

In sociology, Niklas Luhmann introduced the term expectations of expectations (*Erwartungserwartungen*) [51]. This recursivity is integral to understanding us as intentional and cooperative agents. A technological system cannot have

[3] "We thus proposed that the basic cognitive skill of shared intentionality– recursive mindreading– arose as an adaptation for collaborative activity specifically (given an initial adaptation in the direction of tolerance and generosity with food), leading to the creation of joint attention and common ground. The combination of helpfulness and recursive mindreading led to mutual expectations of helpfulness and the Gricean communicative intention as a guide to relevance inferences" [50, p. 217/218].

expectations. Certain autonomous systems might be interpreted as having intentions, but they cannot be understood as cooperative agents. They have certain dispositional features that let them react in a certain way [51]. The system might be *able* to choose means and ends, which could be described as reactions of dispositions they possess. There are specific possible means or ends available and form a realm or space of possibilities that can be expressed in counterfactual statements: "If systems S would have chosen means M, it would have resulted in output O." Dispositions of intelligent systems express if they are reliable and make "good" discriminations or inferences, "*ranges of counterfactual robustness*" [52, p. 103]. We humans are discursive and normative beings. We make commitments, and we can challenge them or give reasons for them in the form of, for instance, principles, theories, and factual statements [47, 52]. We operate in a "space of reasons"[4] [53].

4 User modeling

User expectations can be obtained by collecting data from users, tracking them, gathering information from online reviews [54], and (automatically) inferring their expectations from semantic contexts [55]. Such implicit feedback is located more in a behavioral framing. This is also the case for "behavioral evaluations such as A/B tests measuring user activity instead of surveys or ethnographic analyses that capture users' experience and perception of the system" [56, p. 221]. Explicit feedback can also be obtained via participatory user studies [57] or surveys. Furthermore, user-system negotiations via dialogues with (Generative) AI might present new ways to engage with users [49]. Technological systems can also create models of users, for instance, through generalization. In doing so, the user expectations can be met or disappointed. A movie recommendation system, for example, can meet expectations by being optimized via implicit or explicit feedback [58, 59]. The systems (based on a neural network, for instance) cannot establish expectations but only (weighted) distributions of probabilities. Consequently, in human-system co-actions, no expectations can be aligned like it is possible between humans. However, expectations play a crucial role in the difference between the expected outcome (inputted prompt) and the realized outcome (generated picture, text).

Someone who paints a landscape picture has an idea in her head what she wants to paint.[5] Nevertheless, it depends on the acquired abilities to paint how close one comes to the imagined picture. Generative AI offers here new forms of amplifying or substituting human abilities. Painting a picture can also be accompanied by disappointment. Still, an experienced painter or someone else (a teacher) could at least offer advice to improve the picture and advise on which abilities need to be trained to paint a better picture. However, when using Generative AI, it is mostly not clear what would lead to disappointment; for instance, the disappointment of the expectation could be based on (1) systemic strategies (determined by the developers), for example, for predictions ("synthetic users" [60], user personas), preset rules (affirmative action policies), etc., (2) the coordination via third parties ("stereotypes" [61], "anonymous communitization") [60, 62, p. 153], (3) learning or adaptive algorithms (machine learning, neural networks), (4) an inaccurate use [27], (5) implicit or (6) explicit feedback [58, 59]. A system that generates pictures might have an implemented affirmative action policy that outputs for the prompt "nurse" 50% male nurses and 50% female nurses, even though the system was trained mainly with pictures from female nurses. The clustering of users by some characteristics that they share is an example of the second form, where it is mostly not transparent for the user how the system clusters and which features are crucial to, for example, recommend a specific movie. An inaccurate use might be an incomplete or unspecific prompt. Implicit feedback is obtained by monitoring the user, and explicit feedback is received by likes from the user or, more elaborately, by research (for instance, acceptance studies).[6]

Assessing the systems might now seem hopeless because the algorithms are mostly black boxes in Generative AI models. However, the alignment of prompted input and realized output could also be systematically examined to understand how the system works. Text or pictures must be generated and categorized preliminary into the six types above to examine the system. This needs to be further investigated empirically to describe human-system co-action systematically. If the expectations are disappointed, the types can only be inferred abductively from the effects (output). Empirical research is needed to assess the outputs and to understand under which type the specific output is realized. Some use cases and examples will be shortly discussed in the next section as a first step toward more systematic empirical work.

[4] This is also why choosing adequate means [29, p. 8/9], as dispositional systems do, cannot be manipulative but only if a normative commitment is incorporated.

[5] Generative AI can also help generate ideas in a co-creation process [71]. I am thankful to an anonymous reviewer for indicating this issue.

[6] Generative AI can also be used to improve system transparency by introducing new ways of negotiations between users and systems. (I am thankful to an anonymous reviewer for pointing this out.) This issue is also discussed in [48, 49]. I believe that this is certainly a way to approach the ethical issues surrounding Generative AI. Nevertheless, the aim is here not so much to enhance system transparency but rather to make the practice of developing Generative AI more transparent.

5 Ethical implications– use cases

The use of Generative AI poses ethical challenges: For instance, a prompt like "CEO of a big company" outputs mainly pictures of male managers and "nurse" only female nurses (webpages used: davinci.ai/ and craiyon.com). The conception of fairness is not reflected. This might lead to technology resembling reality (as unjust as it might be). Of course, which measurements should be taken in these cases is also unclear. Should the output be randomized, should an affirmative action policy via a fixed rule be implemented (e.g., to output 60% female managers), or should the output mirror an (unjust) reality? Yet, conceptions of fairness, such as affirmative action policies, are also limited in such contexts. They can lead to the generation of absurd pictures, as has been done recently by the Google program Gemini that outputted indigenous people in German World War Two uniforms [63]. The ethical principles behind such technologies must be made explicit and discussed, not only in the offices of tech companies but also in exchange with the broader public.

In the health sector, prioritizing patients for referral to specialists is crucial. This is partly done by using natural language processing (NLP). However, Generative AI also has innovation potential, where a chatbot might function as an assistant to the physician [64, 65]. The prioritization is based on increasing efficiency and effectiveness. Nevertheless, such systems can also systematically discriminate against minorities, as has been discovered in studies [66]. The normative frame, i.e., here, conceptions of fairness, must be discussed and operationalized; otherwise, the conception of prioritization solely relies on the technological categories of efficiency and effectiveness.

Values such as fairness are abstract in the sense of meaningless if they are not connected with other notions like moral desert, equal shares, and randomization. These connections are commitments that have conceptual content that can be challenged. Awad et al. discuss the example of prioritization for kidney transplants. According to them, prioritization could be formalized in descriptive ethics via studies of the preferences or intuitions of laypeople regarding the case to extract features that matter. Furthermore, they claim that the means of prioritization could be based on the notion of the "general health" of the patients [67, p. 391]. However, the value of fairness is conceptualized via a notion of moral desert, i.e., one deserves a kidney over another person based on a specific feature (may it be acquired through effort (consistent dietary) or purely accidentally, such as age). Other ways of conceptualizing fairness, such as implementing a rule (waiting list ("first come, first served")) or randomization [68], are not discussed. The ethical task is discussing different strategies that can be used and how they are justified. The decision on which strategy is used is, in the end, not ethical anymore but pragmatic-moral within the context of a specific political system and culture. Awad et al. propose a computational approach to ethics that presupposes already a specific conceptual content for the value of fairness [67]. Algorithmizing or formalizing normative ethics means either playing a formal game with well-defined formulas where no justification can be sought or presupposing a specific notion that is not questioned.

Another example of using Generative AI is the personalization of learning environments to generate tailor-made learning paths and exercises for students [69]. In particular, the system strategies for personalization, but also the normative background of the concept, would have to be further questioned since, on the one hand, it would have to be shown in a technically transparent way under which strategies the systems are adapted to the users and, on the other hand, to what extent personalization in education makes sense. Even though it promises a fairer education due to adaption to specific learning needs, personalization– and technology in general [70]– reduce resistances that are essential for learning and are actually needed for the development of competencies or skills.– In all examples, a clear reflection and discussion of the ethical and normative framework is lacking. Participatory research is essential here in establishing how values should be conceptualized and which conceptualization stakeholders consider (ethically) acceptable. Such investigations of the acceptability of technological systems can help to prevent such systems from being manipulative. Although they might influence the users, a deliberative process would have been established to find and choose acceptable conceptualizations.

6 Conclusion

Seemingly, Generative AI could help to make technological systems more transparent because it can provide feedback or interact with users via dialogues. For this reason, it has been proposed to implement a theory of mind in systems. However, it has been shown that this conceptualization has several shortcomings. Already in the developmental phase, it is crucial to align the expectations of developers and users because the user is influenced or manipulated through the mediating role of technology. Even if the developers tried not to intentionally deceive the user via, for instance, dark patterns, some kind of non-intentional manipulation might be the consequence of the system implementation. Hence, in the process of human-system co-action, this kind of influence or manipulation might already be taking place and, furthermore, be opaque to the users and, even worse, not in their interest.

Systems cannot form expectations to conform to users' interests, goals, or values. In the end, it is only possible to align expectations between humans: developers and designers on the one side and users on the other side. The discussed use cases showed that it is crucial to understand, for instance, which conception of fairness is and can be implemented in a system. For this reason, the preferences and values of the stakeholders need to be explored. Which conception of fairness do they favor? And how can we democratize Generative AI?

Much work remains to be done before a full understanding of the ethics of Generative AI is established. In terms of future research, it would be crucial to extend the present discussions by examining the operationalization of ethical values, such as fairness, privacy, etc., in technological systems. Furthermore, it needs to be investigated how systems strategies can be made transparent to users so that users also can find counter-strategies or align their expectations. If it is possible to implement different conceptions of values, then there is a need for empirical research that explores the preferences of the stakeholders and which conception they would favor. This implies converting or transforming human-system co-action to a more transparent human-system interaction framework.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The author has no relevant financial or non-financial interests to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Verbeek, P.-P.: Beyond interaction: a short introduction to mediation theory. Interactions **22**(3), 26–31 (2015)
2. Rosenberger, R., Verbeek, P.-P.: A Postphenomenological field guide. In: Rosenberger, R., Verbeek, P. (eds.) Postphenomenological Investigations: Essays on Human-Technology Relations, pp. 9–41. Hrsg., New York/London (2015)
3. van de Poel, I., Verbeek, P.-P.: Can technology embody values?. In: Kroes, P., Verbeek, P. (eds.) The Moral Status of Technical Artefacts. Philosophy of Engineering and Technology, vol. 17, pp. 103–124, Heidelberg/New York/London (2014)
4. Krügel, S., Ostermaier, A., Uhl, M.: ChatGPT's inconsistent moral advice influences users' judgment. Sci. Rep. **13**, 4569 (2023)
5. Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., Rahwan, I.: The moral machine experiment. Nature **563**(7729), 59–64 (2018)
6. Krügel, S., Uhl, M.: Autonomous vehicles and moral judgments under risk. Transp. Res. Part A: Policy Pract. **155**, 1–10 (2022)
7. Karpus, J., Krüger, A., Tovar Verba, J., Bahrami, B., Deroy, O.: Algorithm exploitation: humans are keen to exploit benevolent AI. iScience **24**(6), 1–16 (2021)
8. Schönmann, M., Bodenschatz, A., Uhl, M., Walkowitz, G.: The care-dependent are less averse to care robots: an empirical comparison of attitudes. 15, (2023). Int. J. Soc. Robot. **15**, 1007–1024 (2023)
9. Hubig, C.: Die Kunst des Möglichen III: Macht der Technik, Bielefeld (2015)
10. Wang, Q., Saha, K., Gregori, E., Joyner, D., Goel, A.: Towards mutual theory of mind in human-AI interaction: how language reflects what students perceive about a virtual teaching assistant. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, pp. 1–14 (2021)
11. Eicher, B., Cunningham, K., Gonzales, S.P.M., Goel, A.: Toward mutual theory of mind as a foundation for co-creation. In: International Conference on Computational Creativity, Co-Creation Workshop (2017)
12. Ropohl, G.: Technikbegriffe zwischen Äquivokation und Reflexion. In: Banse, G., Grunwald A. (eds.) Technik und Kultur: bedingungs- und Beeinflussungsverhältnisse, pp. 41–54, Hrsg., Karlsruhe, (2010)
13. VDI, Richtlinie 3780: Technikbewertung, Begriffe und Grundlagen, Berlin (2000)
14. Pitt, J.C.: "Guns Don't Kill, People Kill"—values in and/or around technologies. In: Peter, K., Verbeek, P. (eds.) The Moral Status of Technical Artefacts. Philosophy of Engineering and Technology, vol. 17, pp. 89–101, Dordrecht (2014)
15. Coeckelbergh, M.: Moved by Machines: Performance Metaphors and Philosophy of Technology, New York (2019)
16. Hubig, C.: Die Kunst des Möglichen I: Technikphilosophie als Reflexion der Medialität, Bielefeld (2006)
17. Grunwald, A., Julliard, Y.: Technik als Reflexionsbegriff: Überlegungen zur semantischen Struktur des Redens über Technik. Philos. Nat. **42**, 127–157 (2005)
18. Moore, J.: Four kinds of ethical robots. Philos. Now **72**, 12–14 (2009)
19. Köbis, N., Bonnefon, J., Rahwan, I.: Bad machines corrupt good morals. Nat. Hum. Behav.Behav. **5**(6), 679–685 (2021)
20 Krügel, S., Ostermaier, A., Uhl, M.: Zombies in the loop? Humans trust untrustworthy AI-advisors for ethical decisions. Philos. Technol. **35**, 17 (2022)
21 Krügel, S., Ostermeier, A., Uhl, M.: Algorithms as partners in crime: a lesson in ethics by design. Comput. Hum. Behav.. Hum. Behav. **138**, 107483 (2023)
22. Leib, M., Köbis, N., Rilke, R., Hagens, M., Irlenbusch, B.: Corrupted by algorithms? how ai-generated and human-written advice shape (dis)honesty. Econ. J. **134**(658), 766–784 (2024)
23. Robinette, P., Li, W., Allen, R., Howard, A., Wagner, A.: Overtrust of robots in emergency evacuation scenarios. In: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 101–108 (2016).

24. Thaler, R.H., Sunstein, C.R.: Nudge: The Final Edition, New Haven/London (2021)
25. Thaler, R.H., Sunstein, C.R.: Nudge: The Final Edition, Penguin Books (2021)
26. Klincewicz, M.: Artificial intelligence as a means to moral enhancement. Stud. Log. Gr. Rhetor. **48**(1), 171–187 (2016)
27. Hubig, C.: Die Kunst des Möglichen II: Ethik der Technik als provisorische Moral, Bielefeld (2007)
28 Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., Floridi, L.: The ethics of algorithms: key problems and solutions. In: Floridi, L. (ed.) Ethics, Governance, and Policies in Artificial Intelligence, pp. 97–123. Springer (2021)
29 Klenk, M.: Ethics of generative AI and manipulation: a design-oriented research agenda. Ethics Inf. Technol. **26**, 9 (2024)
30 Susser, D., Roessler, B., Nissenbaum, H.: Technology, autonomy, and manipulation. Internet Policy Rev. (2019). https://doi.org/10.14763/2019.2.1410
31 Sunstein, C.: The Ethics of Influence: Government in the Age of Behavioral Science. Cambridge University Press (2016)
32. Noggle, R.: Pressure, trickery, and a unified account of manipulation. Am. Philos. Q. **57**(3), 241–252 (2020)
33. Noogle, R.: The ethics of manipulation. In: The Stanford Encyclopedia of Philosophy, Summer 2022 Edition
34. Bergram, K., Djokovic, M., Bezençon, V., Holzer, A.: The digital landscape of nudging: a systematic literature review of empirical research on digital nudges. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI'22), pp. 1–16 (2022)
35. Valta, M., Menzel, J., Maier, C., Pflügner, K., Meier, M., Weitzel, T.: Digital nudging: a systematic literature review and future research directions. In: Proceedings of the 2022 Computers and People Research Conference (SIGMIS-CPR'22), pp. 1–10 (2023)
36 Nie, L., Zhao, Y., Chenglin Li, C., Lu, X., Liu, Y.: Shadows in the interface: a comprehensive study on dark patterns. Proc. ACM Softw. Eng. **1**(FSE), 204–225 (2024)
37 Gray, C.M., Chen, J., Chivukula, S.S., Qu, L.: End user accounts of dark patterns as felt manipulation. Proc. ACM Hum.-Comput. Interact **5**(CSCW2), 1–25 (2021)
38. Sánchez Chamorro, L., Bongard-Blanchy, K., Koenig, V.: Ethical tensions in UX design practice: exploring the fine line between persuasion and manipulation in online interfaces. In: Proceedings of the 2023 ACM Designing Interactive Systems Conference (DIS'23), pp. 2408–2422 (2023)
39. Scott, A.E., Neumann, D., Niess, J., Woźniak., P.W.: Do you mind? User perceptions of machine consciousness. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI'23), pp. 1–19 (2023)
40. Heyselaar, E., Bosse, T.: Linking theory of mind in human-agent interactions to validated evaluations: Can explicit questionnaires measure implicit behaviour?. In: Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents (IVA'21), pp. 120–127 (2021)
41. Luger, E., Sellen, A.: "Like having a really bad PA": the gulf between user expectation and experience of conversational agents. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16). Association for Computing Machinery, pp. 5286–5297 (2016)
42. Zamora, J.: "I'm sorry, dave, I'm afraid I can't do that: chatbot perception and expectations. In: Proceedings of the 5th International Conference on Human Agent Interaction (HAI'17). Association for Computing Machinery, pp. 253–260 (2017)
43. Liao, Q.V., Mas-ud Hussain, M., Chandar, P., Davis, M., Khazaeni, Y., Crasso, M.P., Wang, D., Muller, M., Shami, N.S., Geyer, W.: All work and no play?. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18). Association for Computing Machinery, pp. 1–13 (2018)
44. Kim, K.-J., Lipson, H.: Towards a simple robotic theory of mind. In: Proceedings of the 9th Workshop on Performance Metrics for Intelligent Systems (PerMIS'09), pp. 131–138 (2009)
45. Hegel, F., Krach, S., Kircher, T., Wrede, B., Sagerer, G.: Theory of mind (ToM) on robots: a functional neuroimaging study. In: Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction (HRI'08), pp. 335–342 (2008)
46. Devin, S., Alami, R.: An implemented theory of mind to improve human-robot shared plans execution. In: The Eleventh ACM/IEEE International Conference on Human Robot Interaction (HRI'16), pp. 319–326 (2016)
47 Brandom, R.: Making It Explicit: Reasoning, Representing, and Discursive Commitment. Harvard University Press (1994)
48. Verma, M., Bhambri, S., Kambhampati, S.: Mudit Verma, Siddhant Bhambri, and Subbarao Kambhampati. theory of mind abilities of large language models in human-robot interaction: an illusion?. In: Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI'24), pp. 36–45 (2024)
49. Deshpande, M., Magerko, B.: Embracing embodied social cognition in AI: moving away from computational theory of mind. In: Extended abstracts of the 2024 CHI conference on human factors in computing systems (CHI EA'24), pp. 1–7 (2024)
50 Tomasello, M.: Origins of Human Communication. The MIT Press (2008)
51. Hubig, C.: Verschmelzung von Technik und Leben? Begriffserklärungen an der Schnittstelle von Mensch und technischem System. In: Herzberg, S., Watzka, H. (Eds.) Transhumanismus: Über die Grenzen technischer Selbstverbesserung, pp. 145–160, Hrsg., Berlin/Boston (2020)
52 Brandom, R.: Between Saying and Doing. Oxford University Press (2008)
53. Sellars, W.: Empiricism and the Philosophy of Mind, 4 Hrsg., Cambridge (Mass.)/London (2003)
54. Svikhnushina, E., Placinta, A., Pu, P.: User expectations of conversational chatbots based on online reviews. In: Proceedings of the 2021 ACM Designing Interactive Systems Conference (DIS'21), pp. 1481–1491 (2021)
55. Hotzkow, J.: Automatically inferring and enforcing user expectations. In: Proceedings of the 26th ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2017), pp. 420–423 (2017)
56. Ekstrand, M.D., Willemsen, M.C.: Behaviorism is not enough: better recommendations through listening to users. In: Proceedings of the 10th ACM Conference on Recommender Systems (RecSys'16), pp. 221–224 (2016)
57. Park, S., Lim, Y.-K.: Investigating user expectations on the roles of family-shared AI speakers. In: Proceedings of the 2020 CHI Conference on human factors in computing systems (CHI'20), pp. 1–13 (2020)
58. Jawaheer, G., Szomszor, M., Kostko, P.: Comparison of implicit and explicit feedback from an online music recommendation service. In: Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec'10). Association for Computing Machinery, New York, pp. 47–51 (2010)
59. Zhao, Q., Harper, F.M., Adomaviciu, G., Konstan, J.A.: Explicit or implicit feedback? engagement or satisfaction? a field experiment on machine-learning-based recommender systems. In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing (SAC'18). Association for Computing Machinery, New York, pp. 1331–1340 (2018)
60. Johnson, A., Taatgen, N.: User Modeling. In: Handbook of human factors in Web design, Lawrence Erlbaum Associates, pp. 4244–39 (2005)

61. Rich, E.: Stereotypes and User Modeling. User Models in Dialog Systems. Symbolic Computation (1989)
62. Hubig, C.: Virtualisierung der Technik—Virtualisierung der Lebenswelt. In: Gethmann, C.F.(Ed.) Lebenswelt und Wissenschaft: XXI. Deutscher Kongreß für Philosophie, pp. 146–159, Hrsg., Hamburg (2011)
63. Google pauses AI-generated images of people after ethnicity criticism. The Guardian (2024)
64 Saqib, M., Iftikhar, M., Neha, F., Karishma, F., Mumtaz, H.: Artificial intelligence in critical illness and its impact on patient care: a comprehensive review. Front. Med. (Lausanne) **1**, 1176192 (2023)
65. Pessach, I., Shaked, O., Lipsky, A., Zeevi, A., Lilly, C., Blum, J.: Focusing advanced clinicians on the more critically ill patients using artificial intelligence. Crit. Care Med. **48**(1), 177 (2020)
66. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019)
67. Awad, E., Levine, S., Anderson, M., Anderson, S.L., Conitzer, V., Chrockett, M., Everett, J.A., Evgeniou, T., Gopnik, A., Jamison, J.C., Kim, T.W., Liao, S.M., Meyer, M.N., Mikhail, J., Opoku-Agyemang, K., Schaich Borg, J., Schroeder, J., Sinnott-Armstrong, W., Slavkovik, M., Tenenbaum, J.B.: Computational ethics. Trends Cogn. Sci.Cogn. Sci. **26**(5), 388–405 (2022)
68 Broome, J.: Fairness. In: *Proceedings of the Aristotelian Society, New Series*, vol. 91, pp. 87–101 (1990/1991)
69. Freier, C., Bocklet, T., Helte, A.-K., Hoffmann, F., Hunger, M., Kovács, L., Richter, F., Riedhammer, K., Schmohl, T., Simon, C.: Wie kann videogestütztes Lernen die Erwartungen Studierender und Dozierender erfüllen? Soziale Passagen **15**(2), 631–635 (2023)
70. Ortega y Gasset, J.: Betrachtungen über die Technik. In: Gesammelte Werke, vol. IV, Stuttgart, pp. 7–69 (1978)
71 Wan, Q., Hu, S., Zhang, Y., Wang, P., Wen, B., Lu, Z.: It felt like having a second mind: investigating human-AI co-creativity in prewriting with large language models. Proc. ACM Hum. Comput. Interact. **8**(CSCW1), 1–6 (2024)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

AI and Ethics
https://doi.org/10.1007/s43681-024-00624-1

ORIGINAL RESEARCH

Democratizing value alignment: from authoritarian to democratic AI ethics

Linus Ta-Lun Huang[1,2] · **Gleb Papyshev**[2,3] · **James K. Wong**[3,4]

Received: 11 July 2024 / Accepted: 16 November 2024
© The Author(s) 2024

Abstract
Value alignment is essential for ensuring that AI systems act in ways that are consistent with human values. Existing approaches, such as reinforcement learning with human feedback and constitutional AI, however, exhibit power asymmetries and lack transparency. These "authoritarian" approaches fail to adequately accommodate a broad array of human opinions, raising concerns about whose values are being prioritized. In response, we introduce the Dynamic Value Alignment approach, theoretically grounded in the principles of parallel constraint satisfaction, which models moral reasoning as a dynamic process that balances multiple value principles. Our approach also enhances users' moral and epistemic agency by granting users greater control over the values that influence AI behavior. As a more user-centric, transparent, and participatory framework for AI ethics, our approach not only addresses the democratic deficits inherent in current practices but also ensures that AI systems are flexibly aligned with a diverse array of human values.

Keywords Authoritarian AI ethics · Value alignment · Large language models (LLMs) · Democratization

1 Introduction

In the development of advanced AI systems, addressing value alignment is crucial to ensure these systems operate in ways consistent with human values. Two primary value alignment challenges exist: the first one (Problem 1.0) involves ensuring AI systems understand and act according to human values. The second challenge (Problem 2.0) concerns determining which human values (whose values) should be prioritized for alignment.

Problem 1.0 involves technical difficulties in aligning AI systems to ensure they act according to human intentions and values. One significant challenge is that human values are often abstract, context-dependent, and can vary widely between different cultures and individuals. Translating these nuanced values into precise, actionable guidelines that an AI system can understand and follow is inherently complex. Additionally, AI systems must be equipped to handle conflicting values and make ethical decisions in ambiguous situations. Ensuring that these systems can generalize from limited training data to a wide range of real-world scenarios without misinterpreting or oversimplifying human intentions further complicates the alignment process.

Problem 2.0 concerns determining which human values should be prioritized for alignment, raising fundamental questions about whose values should be encoded into AI systems. This issue is deeply rooted in the pluralistic nature of human societies, where values can vary widely based on cultural, regional, and individual differences. The selection process for these values can often reflect the biases and perspectives of those in positions of power, such as AI developers and corporate stakeholders, potentially marginalizing less dominant voices. The lack of a broad, inclusive, and democratic process for determining these values can lead to AI systems that disproportionately reflect the interests of specific groups, exacerbating existing inequalities and failing to serve the broader public good.

✉ Gleb Papyshev
gleb@ust.hk

1 Division of Humanities, The Hong Kong University of Science and Technology, Hong Kong SAR, China

2 Centre for AI Research, The Hong Kong University of Science and Technology, Hong Kong SAR, China

3 Division of Social Science, The Hong Kong University of Science and Technology, Hong Kong SAR, China

4 Division of Public Policy, The Hong Kong University of Science and Technology, Hong Kong SAR, China

Published online: 02 December 2024

Moreover, the process of determining which values to prioritize is inherently tied to issues of transparency and accountability in AI development. Without clear, open, and participatory mechanisms for value selection, the public cannot effectively scrutinize or influence the values that guide AI behavior. This opacity can foster mistrust and skepticism towards AI technologies, particularly among communities that feel their values are underrepresented or ignored.

While most of the research in AI alignment is concentrated around Problem 1.0, this article provides a potential solution to Problem 2.0 through the Dynamic Value Alignment Approach. The first section of the article highlights the ways in which two prominent value alignment techniques (reinforcement learning with human feedback and Constitutional AI) pose inherent challenges to democratic value alignment. The second section shows how current value alignment approaches are undemocratic in nature, excluding some individuals from acting as epistemic objects. The last section introduces the approach of dynamic value alignment, which can allow users to exercise greater control over the values guiding the behaviors of the AI systems they are interacting with.

2 The value alignment problem

AI companies employ two primary methods for value alignment in advanced AI systems: reinforcement learning with human feedback (RLHF) and Constitutional AI. These approaches aim to address Problem 1.0, which involves ensuring AI systems understand and respect human values. However, a significant power imbalance exists between developers and users regarding the selection of values to be aligned, potentially leading to models that fail to reflect the diverse values and needs of society.

RLHF, prominently used by OpenAI, involves collecting human-written demonstrations of desired output behavior, training a supervised learning baseline, and then collecting human-labeled comparisons of model outputs to train a reward model that predicts human preferences (Ouyang et al. 2022). While RLHF introduces a human-in-the-loop approach to align AI agents' goals with human-expressed ethical values (Kaufmann et al. 2023), achieving a more representative alignment process necessitates the inclusion of diverse stakeholders, which can lead to conflicting values and potentially compromise the model's performance (Conitzer et al. 2024). Furthermore, alignment through RLHF often results in superficial changes to the model's performance (Lee et al. 2024), and in safety-critical domains, low-confidence feedback from people (situations when people may hardly have any consensus) cmay accelerate potential incidents (Chaudhari et al. 2024). Newer techniques involving AI-generated feedback, while more efficient, may inadvertently incorporate non-aligned information into the system's design (Li et al. 2024). The key challenge to democratic governance of AI models lies in the lack of disclosure regarding the RLHF process, hindering public understanding and oversight of the value alignment process (Casper et al. 2023).

Constitutional AI, championed by Anthropic, relies on embedding a predefined set of principles into the AI's decision-making process (Bai et al. 2022). These principles, chosen by developers, guide the model's behavior and ensure alignment with human values (Conitzer et al. 2024). Once established, the principles are encoded into the AI system's architecture, constraining its decision-making process to operate within the boundaries of what is considered acceptable and aligned with human values (Kundu et al. 2023). However, the current process of designing Constitutional AI is inherently undemocratic, as the principles are determined by a selected group of developers without broad public input or consensus, raising concerns about legitimacy and accountability. The technical complexity of AI systems makes it difficult for the public to understand how the principles are implemented and enforced, leading to a concentration of power in the hands of the developers. Furthermore, the rigidity of hardcoded principles may not allow for necessary flexibility as societal values evolve over time, and there are no mechanisms for individualized principle alignment at this stage.

OpenAI's RLHF and Anthropic's Constitutional AI approaches raise significant concerns about democratic legitimacy and accountability in AI development, particularly in addressing Problem 2.0, which involves determining whose values should be prioritized for alignment. The centralized, opaque, and potentially biased nature of these approaches poses risks to the inclusive and responsible development of AI technologies with far-reaching societal impacts.

One of the primary issues with RLHF is the lack of transparency surrounding the process. The specific details of the human feedback and reinforcement signals used to train the AI are not fully disclosed to the public. This lack of transparency is compounded by the concentration of power in the hands of a small group of individuals at OpenAI who design the training process and select the human feedback providers. Similarly, Anthropic's Constitutional AI approach involves encoding certain values, rules, or "constitutional principles" into the AI during training, but the process of determining these principles and embedding them into the AI is also largely opaque and controlled by a small group of researchers at Anthropic, with limited public input or oversight.

Both RLHF and Constitutional AI face the risk of introducing biases into the AI's outputs. In RLHF, the human feedback used in training comes from a limited sample of people, which could lead to the AI reflecting and amplifying certain perspectives while marginalizing others. Constitutional AI faces a similar risk, as the values and principles encoded into the AI may reflect the preferences, biases, and limited perspectives of the researchers involved in the process. For example, the list of principle categories used by Anthropic includes various frameworks such as the Universal Declaration of Human Rights (which has numerous interpretations) and Apple's Terms of Services, yet the reasons for selecting certain frameworks over others remain unclear (Bai et al. 2022). Though there are attempts to democratize Constitutional AI through the engagement of the wider public for constitution drafting, such involvement has been limited to a thousand Americans only, leaving the vast majority of the model's potential users outside of value alignment conversation (Anthropic 2023).

The concentration of power in the hands of a few individuals or organizations in the development of AI systems raises significant concerns about the potential for power disparities to influence the value alignment process. When a small and homogeneous group is responsible for determining the guiding values and principles, there is a risk that the resulting systems will reflect and perpetuate existing power structures and inequalities, potentially marginalizing and excluding already disadvantaged communities.

In the case of OpenAI's RLHF, the selection of human feedback providers is a critical point where power disparities can come into play. If the feedback providers are not representative of the diverse range of perspectives and experiences in society, the resulting AI system may be biased towards the values and preferences of the dominant group, exacerbating existing inequalities. Similarly, in Anthropic's Constitutional AI approach, the determination of the "constitutional principles" that guide the AI's behavior is a key point where power disparities can influence the outcome.

Moreover, the lack of public input and oversight in both RLHF and Constitutional AI approaches means that there are limited opportunities for those outside the dominant power structures to challenge or influence the value alignment process. This can lead to a situation where the values and interests of powerful actors, such as large technology companies or well-funded research institutions, are prioritized over the needs and concerns of the wider public.

3 The ethics of AI ethics

In response to Problem 2.0, it is necessary to delve into the *ethics* of AI ethics, which refers here to the right and wrong of aligning the AI systems (such as large language models, LLMs) with human values. As pointed out, the Problem 1.0 is more formal as it assumes that there is a unity or even consensus of human values to be aligned with by the AI systems. On the other hand, value pluralism suggests that there exists a plurality of values, many of which are non-comparable and non-commensurable. This empirical fact exacerbates the challenge which concerns not only the means but also the end of value alignment, i.e., not only *how* but also *what* values should be aligned. In this way, there is a need to address normative questions in both the means and the end dimensions of value alignment.

Human individuals make moral decisions guided by a variety of moral values. Consider the classic thought experiment of the trolley problem (Foot 1967; Thomson 1976, 1985). A trolley is running down a track and is going to kill five persons at the end of the track. There is a lever on the side, and if it is turned, the trolley can be directed to another track and will kill one person at the end of that track. The question is whether we should turn the lever. It would be normal to expect that different individuals will react to this thought experiment with different value considerations and decisions. A typical utilitarian, for example, would highlight the number of lives that can be saved as a result of switching the track. A typical deontologist, on the other hand, would reject the consequentialist reasoning and question about the rightfulness of intentionally killing an innocent person on the other track. In addition, even among the utilitarians, it is possible that they disagree on how utilities should be calculated as well as whether the focus should be on individual actions (i.e., act utilitarianism) or on general rules and principles (i.e., rule utilitarianism).

While the trolley problem is hypothetical, similar moral dilemmas do exist in the real world. Imagine a self-driving car is in the face of a choice between crashing into a heavy obstacle ahead which will kill five passengers onboard or swerving the car onto the sidewalk which will kill one pedestrian. The self-driving car makes decisions based on the algorithms programmed. Given the background of value pluralism, similar to the trolley problem, we can reasonably expect a plurality of value considerations and decisions. The question of value alignment, particularly Problem 2.0, arises when we need to decide which of the moral values and principles should ever be programmed into the algorithms. The moral decisions of a self-driving car depends on humans' moral choices in value alignment, both with significant moral implications (Gurney 2016; Nyholm and Smids 2016; Contissa et al. 2017; Awad et al. 2018).

When it comes to moral decision-making, it is reasonable to expect that the decision-makers are morally responsible, or accountable to others, for their choices. Though it remains debatable whether the notion of moral responsibility can ever be applicable to AI systems, it is at least plausible to hold human individuals responsible for their moral decisions, such as their responses to moral dilemmas as well as their choices of values and principles for alignment with AI systems. Individual humans are morally responsible for their decisions if and only if they are autonomous. People are autonomous when they are able to make free choices as self-determining individuals. In general, this refers to a situation in which people are not subject to the control of external constraints (e.g., threats or coercions) or internal compulsions (e.g., extreme negative emotions), have capacities to deliberate rationally (e.g., understanding background information, assessing available options, and deriving reasoned conclusions), and they genuinely believe in the relevant values that underpin their decisions. In this way, individuals are said to be self-determining or autonomous, who function as a moral agent capable of bearing moral responsibility. It is worth noting that people can still be regarded as autonomous even when they make moral decisions under some minimal constraints, such as those consistent with human flourishing and/or interdependencies (e.g., substantive and relational autonomy) (Phillips 2021; Nedelsky 2012).

Autonomous agents exercise rationality when making moral decisions. Despite there are numerous conceptions of rationality, moral decision-making can be considered rational only if a certain logic of inference is followed. A common way for inferring moral decisions is through moral reasoning. For this approach, moral judgments on specific cases are derived from some general moral principles and descriptive claims about the cases. For instance, in the trolley problem, the decision of turning the lever could be made based on the following: (1) One should always act in a way that maximizes the overall happiness; (2) turning the lever will divert the train to another track, killing one person while saving five persons; (3) saving five persons and killing one person can generate more happiness than saving one person and killing five persons; and therefore, (4) one should turn the lever. Here, the moral judgement (i.e., statement (4) is derived from the moral principle (i.e., statement (1) and the empirical claims (i.e., statements (2) and (3). An alternative approach would be to follow what it is called moral reflection (Burnor and Raley 2018). Instead of relying on general moral principles, moral judgements are made based on the descriptive claims and moral considerations about the persons and/or interpersonal relationships involved. For example, one would consider the relevant facts about persons involved in the trolley problem as well as the relationships between these persons and the decision makers. This approach is common for the perspectives of virtue ethics and care ethics. Regardless of which approach to moral decision-making is adopted, it is clear that certain logic of inference is present, which is arguably a minimal requirement for rationality.

Whether through moral reasoning or moral reflection, moral agents reach statements or claims that they believe as appropriate or true. These beliefs can be regarded as knowledge if the agent have reasons to believe them to be true. For example, in the trolley problem, one can confirm whether the descriptive claims, i.e., statements (2) and (3), are true through collecting and interpreting empirical evidence. For the moral principle, i.e., statement (1), one can also justify why it is the most foundational principle that should be accepted as the starting point for moral reasoning. All in all, moral agents are possessors of knowledge, be it empirical or moral knowledge, which is essential for rational moral decision-making. In short, epistemic agency is a precondition of moral agency.

Risse (2023) suggests that, in digital worlds, human individuals can be considered epistemic agents who are both knowers and knowns. It means that humans are both epistemic subjects and epistemic objects at the same time, where both roles are relevant for information exchange with digital systems, such as AI systems. On the one hand, epistemic subjects seek and obtain information about the world. On the other hand, epistemic objects reveal and generate information about themselves. This highlights a unique context for moral decision-making in which humans and digital systems interact to form empirical and moral claims and hence moral judgements. As epistemic subjects, human moral agents can obtain information (through AI systems or not) about the moral issue in question as well as the available options and moral principles. As epistemic objects, human moral agents can generate information about how *they* understand the moral issue in question, what available options *they* think there are.- After sharing such information with the AI systems, the moral agents will likely receive feedback information from AI systems that ultimately help shape their moral reasoning or reflection. Back to Problem 2.0. Regarding whose values should AI systems align with, it would be straightforward to resolve the problem via 'authoritarian' approaches which consider the inputs of only a limited number of epistemic agents, such as algorithm developers or human feedback providers.

As pointed out, the background of Problem 2.0 is value pluralism, in which different individuals and cultures hold diverse, conflicting and irreducible values. Undemocratic value alignment excludes the users from acting as full epistemic agents, and as a result, full moral agents. This is problematic in two aspects. First, some human moral agents are excluded from revealing and generating information

that can shape collaborative moral reasoning and reflection when interacting with AI systems. This is similar to the act of silencing, or external exclusion, in public deliberation (Young 2022). Second, even if human moral agents can provide inputs to shape their interaction with the AI systems, the information revealed may not impact the values embedded in the AI system effectively to guide the responses generated. In a sense, the information provided is not treated as credible and/or intelligent and hence given less consideration by AI systems. This is similar to internal exclusion in public deliberation (Young 2022). Such exclusion will result in a form of testimonial injustice (Fricker 2007), in which these human agents may then hold back and silence themselves as in self-censorship (Festenstein 2015). In short, the current value alignment approaches undermine users' full epistemic and moral agency.

4 Dynamic value alignment

To address the problems inherent in the current value alignment approaches, we propose the Dynamic Value Alignment. This approach enables users to exercise greater moral and epistemic agency when interacting with AI systems. As a result, the users can be more responsible and accountable for the decisions made in collaboration with these systems. To help users exercise their agency, this approach aims to (1) give users more real-time control over the diversity of values and their priorities that guide the generation of responses of LLMs, (2) generate diverse response options that are constrained by general moral outlook, (3) provide transparent and interpretable morally relevant information about each of the alternative responses to facilitate informed decision-making. In the following, we first discuss the theoretical foundation behind this approach, parallel constraint satisfaction. Then, we articulate how to implement an approximate model that is tractable and flexible in LLMs through a judgment aggregation algorithm to enhance users' autonomy.

4.1 Moral reasoning as parallel constraint satisfaction

Human reasoning and decision-making can be understood through the lens of constraint satisfaction processes (Simon and Holyoak 2003; Thagard 2000). This framework is particularly useful in explaining complex cognitive tasks that need to meet multiple constraints. Constraint satisfaction process is initially proposed to explain the coherence effect where the human brain, starting from a state of conflict or ambiguity, strives to arrive at a stable understanding or decision that achieves internal consistency among its beliefs, judgments, emotions, etc. To illustrate this effect, imagine an individual brain as a jury deciding whether a defendant is guilty based on conflicting pieces of evidence. Initially, jurors may have contradictory and uncertain opinions as they weigh the diverging pieces of evidence. However, as they deliberate and discuss the case, their conflicting opinions start to resolve. Over time, a coherent narrative emerges among the jurors, wherein the evidence supporting one conclusion (e.g., innocence) becomes more compelling, while contradictory evidence loses its influence. By the end of the deliberation, the jury reaches a consensus verdict.

According to the constraint satisfaction framework, the coherence effect is achieved through the brain finding a state that satisfies the maximum number of constraints. Using a simplified neural network as illustration, different nodes in the network represent different constraints, i.e., different pieces of evidence and jurors. Nodes start out with various levels of activation, where nodes that agree with each other (e.g., supporting a particular verdict) excite each other, while those that disagree inhibit each other. The process of coherence maximization involves parallel, iterative updating of the activation levels of nodes. Each node receives inputs from connected nodes, adjusting its activation based on the sum of excitatory and inhibitory inputs. Ultimately, this iterative process results in a state where the overall activation pattern reflects a final decision that best satisfies overall constraints.

Moral reasoning is a special case of the more general phenomenon of human reasoning, as a result, it is subject to the same general principle of constraint satisfaction despite its unique features (Thagard 1998). In determining what is right or wrong, at least two important types of constraints are involved: (1) various empirical facts (as evidence), including consequences, intentions, characters, previous judgments, social contexts, etc., and (2) different moral frameworks (as jurors), including moral theories, such as utilitarianism and Confucian virtue ethics, as well as moral intuitions, such as care and fairness according to moral foundation theory. Instead of viewing moral judgment as an effort to maximize adherence to a specific moral theory, this perspective acknowledges that individuals often face moral dilemmas where multiple principles and intuitions conflict. The cognitive process resolves such dilemmas through parallel constraint satisfaction processes, where various moral frameworks and intuitions act as constraints to be considered. The moral judgment that emerges is one that optimally satisfies these competing constraints, not one that maximally fulfills any single moral directive. More importantly, this perspective also accommodates the variability of moral judgment: divergent moral judgments of different individuals reflect different configurations of constraints.

4.2 Implementation through judgement aggregation

Our Dynamic Value Alignment approach leverages the principles of constraint satisfaction in moral reasoning to enable users to exercise greater moral and epistemic agency. To make the implementation of constraint satisfaction more tractable, however, we draw on judgment aggregation procedures as an approximate method (List 2013).

Schematically, this approach will treat an AI system as a jury comprising many moral modules, each responsible for evaluating response options based on evidence and its own system of normative principles. Alternative response options can be generated by the LLMs first. Then, they, along with relevant contextual information, are presented to the moral modules for evaluation. Each moral module will generate a numerical score on a response (e.g., higher positive values for responses that align better with its normative principles, and vice versa). Different moral modules will be given different weights, which should reflect the user's own value priorities. The weighted numerical scores from different modules are then aggregated to form a collective evaluation, which represents how well the response satisfies the overall constraints of the user's moral reasoning. This process is repeated for all response options. Note that the aggregated scores for different options are not the view of a single moral theory or intuition but rather a reflection of an individual user's diverse moral constraints. Finally, the high-scoring options are displayed to the user, along with relevant explanations. This further enables the user to make informed decisions in selecting the response that best aligns with their personal values. Through explicit or implicit feedback, the user will also continue to provide information to adjust the system to better reflect their value priorities.

In the following, we articulate the steps to develop a dynamically aligned AI system. The first step involves embedding a wide range of diverse values into individual modules respectively. This can include operationalizing principles of an ethical theory (e.g., Kantian ethics), an individual dimension of a value theory (such as fairness from moral foundation theory), or rule of thumbs for social norms. This could also involve distilling the implicit normative principles behind moral intuitions of specific cases in specific contexts (e.g., best practice in a medical context). The development of these modules might involve Constitutional AI, RLHF, or reward models, depending on their suitability for extracting and representing specific moral principles. The additional benefits of building individual moral modules to represent distinct normative principles are to represent the diverse sources of values, to prevent unintended interference between them, and to enable flexible combination of moral modules in moral reasoning.

The second step focuses on creating an algorithm capable of generating a diverse range of response options. This algorithm might function independently from the moral modules (such as a standalone LLM module) or involve them. However, the key is to ensure the generated responses are diverse but adhere to basic ethical standards (such as helpful, honest, and harmless).

In the third step, we design a process where the generated response options (together with relevant contextual information) are presented to the moral modules for evaluation. Each module will assign a score reflecting the degree to which an option aligns with its values. A method for normalizing these scores will be developed to ensure that no single module unduly dominates the decision-making process.

The fourth step is to design an algorithm that aggregates the scores from various moral modules. One benefit of this algorithm is that the context-sensitivity of the value alignment will result naturally from this process that approximate parallel constraint satisfaction. Also, we can avoid a response that maximizes a particular value to the extreme, which typically leads to undesirable behaviors. Most importantly, this algorithm will allow for flexibility in assigning weights to different modules, adapting to diverse user value priorities and contexts. The weighting system could be either determined explicitly by the user, or learned implicitly through user interactions, using a reinforcement learning approach, for example. The ability for the user to influence the weights of different moral modules provides a more transparent process for them to control value alignment of the AI system. This stands in stark contrast to the opaque "authoritarian" value alignment approaches discussed earlier.

Developing a user interface constitutes the fifth and final step. This interface will display the high-scoring options to the users, along with explanations regarding some of the most relevant moral principles supporting or undermining them as well as relevant empirical facts. What moral principles contribute most significantly are explicitly represented in the system as the evaluation scores of and weights assigned to the modules. The explanations can be generated by the respective moral modules. With such information, the user can then make a more informed decision in selecting the response that best aligns with their personal value priorities. Importantly, this process contributes to the necessary "knowledge conditions" of responsibility for their decisions, which refers to the requirement that an individual must have adequate awareness and understanding of the relevant facts and circumstances surrounding their actions to be held morally responsible for them (Fischer and Ravizza 2000).

Another important role of the interface is to include a mechanism for users to provide feedback about their value priorities either explicitly or implicitly (e.g., inferred from

their choices). This information is then used to adjust the weights assigned to different moral modules, enabling dynamic adjustment to individual users. This feedback mechanism is crucial because it enables users to exercise their epistemic agency while interacting with the AI system: the users' revealed value priorities have the causal power to significantly shape their collaborative moral reasoning with the AI system.

5 Conclusion

We have discussed the importance of value alignment in developing AI systems to ensure they operate in a way that aligns with human values. We have suggested two challenges for value alignment: Problem 1.0 involves technical difficulties in aligning AI systems with human intentions and values, while Problem 2.0 concerns deciding which human values to prioritize for alignment. Conventional approaches like reinforcement RLHF and Constitutional AI face issues of power asymmetry and lack of transparency in determining values for alignment. These approaches risk excluding diverse values and introducing biases into AI systems. Additionally, they also undermine the moral and epistemic agency of users while interacting with AI systems.

In response to these deficiencies, we proposed Dynamic Value Alignment, which advocates for a paradigm shift towards a more user-centric approach. It respects the complexity of human moral reasoning, where diverse moral and empirical constraints are considered simultaneously. This approach allows users to have real-time control over the diversity and priority of values governing AI behaviors, thereby enhancing users' autonomy and responsibility for their collaborative decision-making with the AI systems. Our approach encourages a participatory ethos in AI ethics, one that is crucial for ensuring that emerging technologies are aligned with and adaptable to the broad spectrum of individual human values.

Funding No funding was received for conducting this study.
Open access funding provided by Hong Kong University of Science and Technology

Data availability Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

6. References

Anthropic: *Collective Constitutional AI: Aligning a Language Model with Public Input*. (2023). https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J., Rahwan, I.: The Moral Machine Experiment. Nature. **563**, 59–64 (2018)

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kaplan, J.: *Constitutional AI: Harmlessness from AI Feedback* (arXiv:2212.08073). arXiv. (2022). https://doi.org/10.48550/arXiv.2212.08073

Burnor, R., Raley, Y.: *Ethical Choices: An Introduction to Moral Philosophy with Cases* (2nd edition). Oxford: Oxford University Press. (2018)

Casper, S., Davies, X., Shi, C., Gilbert, T.K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., Hadfield-Menell, D.: *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback* (arXiv:2307.15217). arXiv. (2023). https://doi.org/10.48550/arXiv.2307.15217

Chaudhari, S., Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., Deshpande, A., da Silva, B.C.: *RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs*(arXiv:2404.08555). arXiv. (2024). https://doi.org/10.48550/arXiv.2404.08555

Conitzer, V., Freedman, R., Heitzig, J., Holliday, W.H., Jacobs, B.M., Lambert, N., Mossé, M., Pacuit, E., Russell, S., Schoelkopf, H., Tewolde, E., Zwicker, W.S.: *Social Choice for AI Alignment: Dealing with Diverse Human Feedback* (arXiv:2404.10271). arXiv. (2024). https://doi.org/10.48550/arXiv.2404.10271

Contissa, C., Lagioia, F., Sartor, G.: The ethical knob: ethically-customisable automated vehicles and the law. Artif. Intell. Law. **25**(3), 365–378 (2017)

Festenstein, M.: Self-censorship for democrats. Eur. J. Political Theory. **17**(3), 324–342 (2015)

Fischer, J.M., Ravizza, M.: Responsibility and control: A theory of moral responsibility. Cambridge University Press (2000)

Foot, P.: The problem of abortion and the doctrine of double effect. Oxf. Rev. **5**, 5–15 (1967)

Fricker, M.: Epistemic Injustice: Power and the Ethics of Knowing. Oxford University Press, Oxford (2007)

Gurney, J.: Crashing into the unknown: an examination of crash-optimization algorithms through the two lanes of ethics and law. Albany Law Rev. **79**(1), 183–267 (2016). https://doi.org/10.48550/arXiv.2312.14925

Kaufmann, T., Weng, P., Bengs, V., Hüllermeier, E.: A survey of reinforcement learning from human feedback. arXiv. (2023).

Kundu, S., Bai, Y., Kadavath, S., Askell, A., Callahan, A., Chen, A., Goldie, A., Balwit, A., Mirhoseini, A., McLean, B., Olsson, C.,

Evraets, C., Tran-Johnson, E., Durmus, E., Perez, E., Kernion, J., Kerr, J., Ndousse, K., Nguyen, K., Kaplan, J.: *Specific versus General Principles for Constitutional AI* (arXiv:2310.13798). arXiv. (2023). https://doi.org/10.48550/arXiv.2310.13798

Lee, K., Hwang, D., Park, S., Jang, Y., Lee, M.: *Reinforcement Learning from Reflective Feedback (RLRF): Aligning and Improving LLMs via Fine-Grained Self-Reflection* (arXiv:2403.14238). arXiv. (2024). https://doi.org/10.48550/arXiv.2403.14238

Li, A., Xiao, Q., Cao, P., Tang, J., Yuan, Y., Zhao, Z., Chen, X., Zhang, L., Li, X., Yang, K., Guo, W., Gan, Y., Yu, X., Wang, D., Shan, Y.: *HRLAIF: Improvements in Helpfulness and Harmlessness in Open-domain Reinforcement Learning From AI Feedback* (arXiv:2403.08309). arXiv. (2024). https://doi.org/10.48550/arXiv.2403.08309

List, C.: Social Choice Theory. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Winter 2013). Metaphysics Research Lab, Stanford University. (2013). https://plato.stanford.edu/archives/win2013/entries/social-choice/

Nedelsky, J.: Law's Relations: A Relational Theory of Self, Autonomy, and Law. Oxford University Press, Oxford (2013)

Nyholm, S., Smids, J.: The Ethics of Accident-algorithms for Self-driving cars: An Applied Trolley Problem? Ethical Theory Moral. Pract. **19**(5), 1275–1289 (2016)

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: *Training language models to follow instructions with human feedback* (arXiv:2203.02155). arXiv. (2022). https://doi.org/10.48550/arXiv.2203.02155

Phillips, A.: Unconditional Equals. Princeton University Press, Princeton (2021)

Risse, M.: Political Theory of the Digital Age: Where Artificial Intelligence Might Take Us. Cambridge University Press, Cambridge (2023)

Simon, D., Holyoak, K.J.: Structural Dynamics of Cognition: From consistency theories to constraint satisfaction. In: The Dynamic Perspective in Personality and Social Psychology. Psychology (2003)

Thagard, P.: Ethical coherence. Philosophical Psychol. **11**(4), 405–422 (1998). https://doi.org/10.1080/09515089808573270

Thagard, P.: Coherence in Thought and Action. MIT Press (2000)

Thomson, J.: Killing, letting die, and the trolley problem. Monist. **59**(2), 204–217 (1976)

Thomson, J.: The Trolley Problem. Yale Law J. **94**, 1395–1415 (1985)

Young, I.: Inclusion and Democracy. Oxford University Press, Oxford (2002)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

AI and Ethics
https://doi.org/10.1007/s43681-024-00627-y

ORIGINAL RESEARCH

The prospects for digital democracy

Ivan Mladenović[1]

Received: 30 June 2024 / Accepted: 16 November 2024
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2024

Abstract
This paper aims to answer a basic question: is it possible to forge democratic citizenship through various online tools that are already available? To answer this question, I introduce the conception of *digital political identities*, i.e., the ways in which online environments contribute to creating, maintaining, and changing political identities. Because the well-functioning of democracy rests on citizens with the ability to make informed decisions, vote, and engage in public deliberation, this paper is looking for new and innovative online tools for participating in meaningful online deliberation, acquiring accurate information in the digital space, and making informed voting decisions. By introducing the conception of digital political identities and linking it to online tools that can improve democracy and citizen engagement, I aim to make further progress in cutting edge research on the relationship between digital technologies and democracy. In a nutshell, I am mainly concerned with proposing and defending a normative framework for the use of various online tools that could foster digital democracy.

Keywords Digital political identities · Artificial intelligence · Democracy · Online deliberation · Digital citizenship

1 Introduction

This paper aims to answer a basic question: is it possible to forge democratic citizenship through various online tools that are already available? To answer this question, I introduce the conception of *digital political identities*, i.e., the ways in which online environments contribute to creating, maintaining, and changing political identities. Because the well-functioning of democracy rests on citizens with the ability to make informed decisions, vote, and engage in public deliberation, this paper is looking for new and innovative online tools for participating in meaningful online deliberation, acquiring accurate information in the digital space, and making informed voting decisions. By introducing the conception of digital political identities and linking it to online tools that can improve democracy and citizen engagement, I aim to make further progress in cutting edge research on the relationship between digital technologies and democracy [1, 2].

Proposals about future directions for digital democracy are rooted in theories of deliberative democracy and epistemic democracy. According to the theories of deliberative democracy, the most basic mechanism of democratic decision-making is public deliberation in which citizens can participate as free equals. Proponents of deliberative democracy thus think that democracy should be more open to public discussion in addition to voting. However, more expansive participation by citizens does not necessarily guarantee that better decisions will be made. For that reason, the theories of epistemic democracy address the conditions under which we might expect democratic decision-making processes to lead to good decisions. There is significant overlap between theories of deliberative and epistemic democracy; after all, both come to the conclusion that public deliberation is a crucial decision-making mechanism that contributes to fair and good decisions. Nevertheless, the theories of epistemic democracy consider a specific set of questions about what contributes to knowledge, truth, accuracy, and so on—the necessary elements to ensure that democratic mechanisms can yield good decisions.

In this paper I focus on tools for online deliberation and the conditions under which they could contribute to good outcomes in an epistemic sense. The paper pays special attention to the recent use of artificial intelligence (AI) tools

✉ Ivan Mladenović
ivan.mladenovic@f.bg.ac.rs

[1] Faculty of Philosophy, University of Belgrade, Belgrade, Serbia

Published online: 27 November 2024

in the context of online public consultation and deliberation. In a nutshell, I am mainly concerned with proposing and defending a normative framework for the use of various online tools that could foster digital democracy. The paper is structured as follows. Section 2 examines social science experiments with online deliberation to determine how these experiments might contribute to (more) digital democracy. Section 3 examines current applications of AI tools in democratic processes and assess possibilities for their future use. Section 4 presents an integrated approach to digital citizenship, i.e., a normative framework for meaningful democratic participation online. Section 5 concludes the paper.

2 Online deliberation: from normative Ideal to practice and back again

When it comes to digital technologies, the most widespread democratic innovation is online deliberation. It is both a subject of study in various experiments within the social sciences and part of the practical application of digital tools for public consultation and deliberation occurring from local to global levels. At city and state levels, there are already many digital platforms that enable online consultation and deliberation [3]. Moreover, tools for online deliberation have been used in crowdsourced constitution-making (such as in Iceland and Ireland). Various forms of online deliberation have also been used at the European Union level (such as for the Conference on the Future of Europe) [4].

This section reviews experiments in online deliberation and the conceptions of digital democracy that emerged from them. In the literature on online deliberation, it is now common to differentiate between three domains of research: institutional design conducive to online public deliberation, rules for deliberative processes, and the outcomes of deliberation itself [5, 6]. In their overview of experiments in online deliberation conducted until 2015, Fries and Eilders [5] noted that most had been devoted to the first two domains, mainly treating the items in the first as independent variables and those in the second as dependent variables. However, a more recent overview of the research field by Strandberg and Grönlund [6] shows that the focus of experiments in online deliberation has shifted to items in the third domain. In these more recent experiments the outcomes of deliberative processes are addressed mainly in relation to institutional design.

To better understand these experiments, some clarity is needed on rules for institutional design and deliberative processes. Rules for institutional design usually determine whether online deliberation is synchronous or asynchronous (extended in time over several instances of deliberation), anonymous or non-anonymous, with or without a facilitator, and so on. Rules for deliberative processes typically address equality, reciprocity, civility, the exchange of reasons, and so on. After testing combinations of synchronous and asynchronous design with anonymous and non-anonymous deliberation, Strandberg and Berg [7] conclude that asynchronous deliberation can yield better epistemic results. They also find, somewhat surprisingly, that anonymous public deliberation does not necessarily have detrimental epistemic effects (although this could be the result of other process-related rules). Given their own experiments and the Deliberative Polling experiments, Strandberg and Grönlund recently pointed out that the most important outcomes of online deliberation are related to "increased knowledge among participants" and a reduction in group polarization [6, p. 372]. So, online deliberation studies show outcomes of significant knowledge gain and positive epistemic effects in general.

Deliberative Polling is one of the most important experiments in both online and offline public deliberation [8]. The experiment consisted of public deliberation over the course of a weekend that unfolded among a group of citizens (chosen by representative random sampling) in smaller sessions and with experts in plenary sessions. The sessions were moderated by trained facilitators. At the beginning of the experiment, citizens received balanced material related to the topic of discussion. At the beginning and at the end of the experiment, participants filled out the same questionnaire. As indicated above, online deliberative polls showed positive epistemic effects mostly around knowledge gains and the acquisition of new information. These effects are well documented by the very same experiments conducted in offline settings.

To locate the proper place of online deliberation based on Deliberative Polling, James Fishkin identifies four types of online public consultation [9]. To begin, he distinguishes between nondeliberative and deliberative forms, as well as representative and unrepresentative practices, of online public consultation, all of which have their analogues in the offline world [9, p. 31]. The first type is the nondeliberative and unrepresentative form of public consultation characteristic of self-selected listener opinion polls (SLOPS). SLOPS are nondeliberative because participants express their preferences or vote without the possibility of deliberation; they are unrepresentative because participants are self-selected. The second type is the nondeliberative yet representative form of public consultation characteristic of the conventional opinion poll.

The third type is the deliberative and unrepresentative form of online discussion among self-selected participants, which Fishkin notes is particularly prone to group polarization [9, p. 33]. The fourth and final type is the deliberative

representative form of public consultation exemplified by online Deliberative Polling. Online deliberative polls, Fishkin says, "offer the promise of greater convenience and continuing dialogue" and because "new technologies allow us to experiment with improvements…eventually, we will be able to periodically take the 'pulse of democracy' in a more deliberative manner" [9, p. 29].

In recent times, Landemore argued "for renovating our paradigm of democracy to make the most of the technological opportunities offered by the digital revolution" [10, p. 65]. She defends the conception of *open democracy* based on the principles of deliberative democracy and equal representation. In Landemore's view, open democracy means that democracy is more open to citizens, that it promotes their open-mindedness, and that it is an open-ended process [10]. One of the most pertinent features of open democracy in the present context is openness to new digital technologies. However, Landemore also holds that despite containing the element of representation, open democracy need not be representative democracy. Instead, she envisages "lottocratic" representation (by stratified random selection) and self-selection as more appropriate forms of representation for deliberative participation (note that in thinking citizens can represent other people even when self-selected, Landemore applies a broader notion of representation than Fishkin).

Both forms of representation are linked to deliberative mini-publics in which each citizen has an equal opportunity to participate. Building on the empirical research into deliberative polls and other mini-publics, Landemore argues that new digital tools could contribute to open democracy by enlarging the scale of citizen participation through the multiplicity of mini-publics. In the context of an ideal of open democracy, she describes the open mini-public as "a deliberative unit that is uniquely possible in a digital world" [10, p. 81]. Landemore believes that this is how digital technologies could contribute significantly to a more deliberative democracy in which citizens participate as free equals.

While Landemore's proposal is grounded in the theory of deliberative democracy, Poblet et al. [2] develop the idea of digital democracy in the direction of epistemic democracy (although their preferred conception is also rooted in deliberative democracy). Their main argument is that much of the information, digital tools, and civic technologies already available in the digital space are dispersed and disconnected. As a remedy for such fragmentation of data and digital tools, they propose the conception of *linked democracy* to empower citizen participation. For them, "linked democracy is about turning this plurality into a participatory ecosystem where data, information, and knowledge are connected and shared" [2, p. 28].

The conception of linked democracy centers on linked open data, linked platforms, and linked ecosystems. Although Poblet et al. recognize many contributions toward open data that are relevant to democratic decision-making, they emphasize the need to make these data interconnected and ready for use at scale. Furthermore, it is not just that data should be somehow connected; rather, connections should be forged between various digital tools and platforms for civic engagement within a digital ecosystem.

While Poblet et al. consider the use of various digital tools at local and state levels as examples of linked ecosystems, I find it consistent with their view to also consider the linked ecosystem at a global level. I also believe that recent developments in AI tools can contribute significantly to each of the dimensions of linked democracy.

3 AI tools for digital democracy

The most recent innovation in digital democracy involves fostering online public consultation and deliberation through generative AI tools. AI-empowered public consultation and deliberation opens up new frontiers for renovating democracy, but the use of such tools also brings about new risks in democratic societies. Landemore warns that "deploying AI in democracy has its risks—like data bias, privacy concerns, potential for surveillance, and legal challenges"—but she also points out that "despite its limitations and risks, AI has the potential to bring about a better, more inclusive version of democracy, one that would in turn equip governments with the legitimacy and knowledge to oversee AI development" [11, p. 14]. More specifically, "AI has the potential to usher in a more inclusive, participatory, and deliberative form of democracy, including at the global scale" [11, p. 12]. This section explores some uses of generative AI in the context of online democracy to assess their potential to foster democratic processes. So far, some of the primary applications of generative AI in this context include summarizing the content of online public deliberation, reducing information overload, and problem-solving. Below, I address each application in turn.

Generative AI tools have proven very effective at *summarizing the content of online public consultation and deliberation*. It is important that public deliberation does not become too dispersed and citizens retain a clear understanding of the purpose of their participation and debates. Such clarity could also be helpful for integrating citizen input into policy-making and law-making processes. Thus far, generative AI tools have been used to summarize lengthy deliberation unfolding in text (e.g., comments on certain proposals on civic tech platforms) as well as video-recorded material from online deliberations.

Some of the most widely used platforms for online democratic participation, such as Your Priorities (created by the Citizens Foundation) and Go Vocal platform (formerly CitizenLab platform), recently introduced AI tools to summarize the content of online public consultation and deliberation.[1] In a recent interview, Landemore elaborates on the use of the AI tool Panoramic (created by Make.org) in the French Citizens' Convention on the End of Life [12]. The tool summarized the content of online deliberation from video materials, but Landemore notes a limit in that summary was offered only for plenary sessions involving mostly the opinions of experts. In any case, generative AI tools carry the potential to summarize content from many different deliberative processes—including discussions in smaller groups, the new frontier for their use in online deliberation.

Summation of the content of online deliberation can be seen as a way to *reduce information overload*. However, civic tech platforms also contain many inputs on various topics, proposals, and so on, that could further contribute to information overload. The notion of information overload is thus not limited to online deliberation. Arana-Catania et al. address the problem of this broader view of information overload through the use of Natural Language Processing (NLP) and machine learning tools to create an enhanced version of the Consul platform [13].[2] Although this online tool has many applications, their case study involved an enhanced version of Consul applied to the platform Decide Madrid. Using NLP tools, they showed how the enhanced version of Consul could contribute to reducing information overload by clustering and categorizing proposals, clustering different views and opinions, and summarizing comments on proposals.

Arana-Catania et al. further experimented with requiring participants to use both an enhanced and a standard version of the Decide Madrid platform. Results showed "a reduction in mean time of 40.9% when using the enhanced version" [13, p. 16]. Overall, the study concludes that by reducing information overload, "NLP and machine learning techniques…improved the effectiveness of citizen participation and collective intelligence processes in a significant way" [13, p. 16]. AI-powered functions to reduce information overload as described above are now integral to several civic tech platforms.

One promising future use of generative AI for democratic purposes emerges within the context of *problem-solving*. In 2024, Citizens Foundation launched the AI-empowered tool Policy Synth, which offers possible solutions when presented with a problem.[3] The idea behind this tool is to enable smart crowdsourcing within problem-solving processes. When given a problem statement, Policy Synth uses Large Language Models (GPT-4 and GPT-3.5) to search for the root causes of the problem and subvert them to subproblems. The tool evolves possible solutions and offers pros and cons for each solution. Once the process of searching for solutions is complete, Policy Synth can also recommend policies.

In addition to these primary uses of generative AI, the AI tool was also partly integrated into the Stanford Online Deliberation Platform that Fishkin and colleagues recently used in their online deliberative polls. This self-moderating platform carries the promise of organizing online deliberation at scale. Studies show that the self-moderating platform can facilitate online deliberation equally as well as trained human moderators.[4] Through comparison of the results of in-person deliberative polls (held in Japan in 2012 and 2014) and an online deliberative poll (held in Japan in 2020), Gelauff et al. conclude that "the Platform for Online Deliberation was able to facilitate a set of deliberations comparable to an in-person experience" [14]. A 2022 experiment with the Stanford Online Deliberation Platform conducted by Meta, which included 6,000 participants from 32 countries, further revealed its ability to facilitate online deliberation at a global level [11]. So, one important further application of AI is *facilitating processes of online deliberation*.

The application of AI to date demonstrates the utility of AI tools for online democratic processes. However, this is not to deny the risks usually invoked (some of which were quoted at the beginning of this section) by advanced AI tools and their further applications. While most of the discussions focus on risks at the level of the individual or democratic society as a whole, the remainder of this section addresses some risks specific to online democratic participation. My main point is that as AI tools advance, their inclusion in online democratic processes should be guided by the idea of enabling individuals to forge democratic citizenship through meaningful participation.

So far, the use of generative AI tools in democratic processes and experiments has been limited owing to the current

[1] For the Your Priorities platform, see: https://www.citizens.is/your-priorities-features-overview/. For the Go Vocal platform, see: https://www.govocal.com/platform-online-engagement-toolbox.

[2] For the Consul Democracy platform, see: https://consuldemocracy.org/features/.

[3] For the Policy Synth toolbox, see: https://github.com/CitizensFoundation/policy-synth. The description of Policy Synth is based on the following unpublished conference paper: Bjarnason, R., Gambrell, D., Lanthier-Welch, J.: Using artificial intelligence to accelerate collective intelligence: Policy synth and smarter crowdsourcing. Final submission for ACM collective intelligence conference (2024).

[4] It is noteworthy that additional evidence from experiments is needed in order to further corroborate this statement. I would like to thank an anonymous reviewer for suggesting me that there is need for additional evidence concerning the use of self-moderating platform, especially at a global level.

stage of their development. But as the AI technologies progress, possibilities for their application in democratic contexts will soon expand. It is thus important that the future use of such technologies does not undermine the democratic process itself. One of the main points of democratic citizenship is that through participation, citizens gain new knowledge of not only what contributes to problem-solving and good decision-making but also the entire democratic decision-making process. Still, overreliance on advanced AI tools could contribute to less engagement by citizens and consequently less knowledge about how to meaningfully participate in democratic processes.[5]

This means that it is critical that further applications of AI tools do not undermine citizen capacity for meaningful democratic participation. Rather, they should foster it. For that reason, the next section presents *an integrated approach to digital citizenship* as a possible framework for advancements in digital democracy and the application of AI tools in online democratic settings.

4 An integrated approach to digital citizenship

The aforementioned views of open democracy and linked democracy hold that normative considerations for digital democracy should be grounded in theories of deliberative and epistemic democracy. Although I agree with this view, I argue that there is a need for an even wider perspective when considering digital political identities. In the near future, political identities will be forged entirely through early exposure to digital technologies. This means that education for digital citizenship is necessary from early on to ensure that meaningful online democratic participation is possible.

In other words, the formation and development of political identity in digital citizenship should run parallel to the formation and development of other digital political identities. The argument for this kind of education is based on reasoning analogous to Rawlsian arguments about the compatibility of a political conception of justice acceptable to all citizens and various comprehensive doctrines [15]. But rather than thinking about the principles of political conception at the core of democratic citizenship, the focus here is on the ways in which digital citizenship and meaningful online democratic participation become possible.

For this reason, I defend *an integrated approach to digital citizenship* that consists of two components: *democratic participation* and *education*. While the democratic participation component includes many of the insights from open and linked democracy, I argue that the educational component should also be taken into account when thinking about the prospects for digital democracy. Democratic participation is crucially important to my integrated approach to digital citizenship. It consists of three elements: (1) participating in meaningful online deliberation, (2) acquiring accurate information in the digital space, and (3) making informed (online) voting decisions. My approach integrates the perspectives of deliberative democracy, epistemic democracy, and theories of voting. At the same time, it is also integrative in the sense that all democratic participation elements are included in the educational component.[6]

Much of this paper has focused on online deliberation. However, access to accurate information is both necessary for meaningful online deliberation and important for informed voting decisions. While digital technologies could contribute to this task, online environments are nonetheless conducive to the spread of misinformation and disinformation that could undermine democratic processes [16, 17]. Watson et al. [18] recently argued for the recognition of epistemic rights as human rights to prevent the harms caused by the spread of misinformation.

The research in the social sciences have already identified many possible tools for acquiring accurate information online and reducing the effects of misinformation. Some of these are already in use on online platforms. Rather than comprehensively reviewing online tools, then, I shall focus on the general unifying framework offered by Lorenz-Spreen et al. [19] that integrates various insights from psychology and experimental economics. To wit, their approach includes various nudging and boosting online tools that can help individuals cope with misinformation and search for accurate information in a digital space. The primary difference between nudging and boosting techniques is that the former focuses on choice architecture and the arrangement of online environments to become more attuned to epistemic cues, while the latter focuses on building individual capacities to search (in a self-directed way) for epistemic cues that promote truth online.

[5] There is a genuine worry (suggested by an anonymous reviewer) that focus on the majority view in summaries of public deliberation, may lead to less engagement by participants when their contribution is not visible. However, I think that there are no technological obstacles to use AI tools in order to offer summaries of different points of views. In fact, some of the civic tech platforms already use this functionality.

[6] To be sure, the idea that both participation and education are crucially important for democracy is not new. Its origins can be traced back to J. S. Mill's *Considerations on Representative Government*. What is the novelty of an integrated approach to digital citizenship? First of all, its novelty is in applying this idea specifically to digital space and digital democracy. Moreover, it is different from Mill's view in that education is not provided *only* through participation in democratic processes, but also through formal educational programs. I would like to thank an anonymous reviewer for prompting me to specify the novelty of an integrated approach to digital citizenship.

To illustrate, nudging tools could involve warnings when information is based on unidentifiable sources or sources that are insufficiently accurate, educative nudges to search for references and multiple sources on the subject of an online article, or the creation of friction to check whether information is accurate before it is shared with other people [19, p. 1106]. Examples of boosting techniques include visualizations of the whole information cascade (including its source and history of sharing), pop-up fast-and-frugal decision trees that help individuals to determine the reliability of a source of information, and the possibility for users to completely customize their news feed [19, p. 1106]. What these all show is that nudging and boosting techniques for promoting truth online should be directed toward both online content and information sources.

Online voting is replete with a distinctive set of problems, which is why the word "online" is bracketed in the third element of the democratic participation component of the integrated approach to digital citizenship. In the short term, voting will mostly remain an offline decision-making procedure. One of the major problems with online voting is that it is still not considered sufficiently secure for use in general elections. But even if online voting is not used solely for that purpose in the near future, it is worth noting that the digital democracy tools discussed thus far became significant to informed voting decisions in offline settings. In the medium and long term, they will be even more important as progress is made to address security issues in online voting. One of the advantages of my integrated approach to digital citizenship is that education about online voting could enable citizens to routinely vote online once this decision-making mechanism is secure enough for broader implementation.

5 Conclusion

In this paper, I considered various tools that are already available that could help to forge digital citizenship and foster digital democracy. I focused on very recent tools that apply generative AI within processes of online public consultation and deliberation. With these tools and their future development in mind, I proposed an integrated approach to digital citizenship as a normative framework when thinking about the prospects for digital democracy.

My argument for an integrated approach to digital citizenship unfolded in two stages. First, I introduced the concept of digital political identities and argued that online environments should contribute to forging digital citizenship through democratic participation and education. Second, I argued that an integrated approach to digital citizenship should include digital tools for participating in meaningful online deliberation, accessing accurate information in the digital space, and making informed voting decisions (including online voting, once it is considered sufficiently secure). I argued that this approach is integrated not only because it includes all of these elements but also because the very same elements should be incorporated into education for digital citizenship.

Acknowledgements This paper was presented at the Paris Conference on AI & Digital Ethics, held on June 6–7, 2024, at the Sciences Po. Earlier version of the paper was presented at the conference Ethics of Socially Disruptive Technologies (ESDIT2022), held on October 6–7, 2022, in Oegstgeest, Leiden. I am grateful to participants of both conferences for very helpful discussions.

Declarations

Conflict of interest The author has no competing interests to declare that are relevant to the content of this article.

References

1. Bernholz, L., Landemore, H., Reich, R. (eds.): Digital Technology and Democratic Theory. The University of Chicago Press, Chicago and London (2021)
2. Poblet, M., Casanovas, P., Rodríguez-Doncel, V.: Linked Democracy: Foundations, Tools, and Applications. Springer, Cham (2019)
3. Simon, J., Bass, T., Boelman, V., Mulgan, G.: Digital Democracy: The Tools Transforming Political Engagement. Nesta, London (2017)
4. European Union. Conference on the Future of Europe: Report on the Final Outcome (2022)
5. Friess, D., Eilders, C.: A systematic review of online deliberation research. Policy Internet. **7**(3), 319–339 (2015). https://doi.org/10.1002/poi3.95
6. Strandberg, K., Grönlund, K.: Online deliberation. In: Bächtiger, A., Dryzek, J.S., Mansbridge, J., Warren, M. (eds.) The Oxford Handbook of Deliberative Democracy, pp. 366–377. Oxford University Press, Oxford (2018)
7. Strandberg, K., Berg, J.: Impact of temporality and identifiability in online deliberations on discussion quality: an experimental study. Javnost - Public **22**(2), 164–180 (2015). https://doi.org/10.1080/13183222.2015.1041230
8. Fishkin, J.: Deliberative polling. In: Bächtiger, A., Dryzek, J.S., Mansbridge, J., Warren, M. (eds.) The Oxford Handbook of Deliberative Democracy, pp. 316–329. Oxford University Press, Oxford (2018)
9. Fishkin, J.S.: Virtual public consultation: prospects for internet deliberative democracy. In: Davies, T., Gangadharan, S.P. (eds.) Online Deliberation: Design, Research, and Practice, pp. 23–35. CSLI, Stanford (2009)
10. Landemore, H.: Open democracy and digital technologies. In: Bernholz, L., Landemore, H., Reich, R. (eds.) Digital Technology and Democratic Theory, pp. 65–89. The University of Chicago Press, Chicago and London (2021)
11. Landemore, H.: Fostering more inclusive democracy with AI. Finance Dev. **60**(4), 12–14 (2023)
12. Giesen, L.: As long as citizens' conventions are black boxes, they lack legitimacy Interview with Hélène Landemore and David Mas (Part 1). Democracy Technologies. https://democracy-tech

nologies.org/participation/ai-citizens-convention-france/ (2024). Accessed 1 Jun 2024
13. Arana-Catania, M., et al.: Citizen participation and machine learning for a better democracy. Digit. Gov. Res. Pract. **2**(3), 1–22 (2021). https://doi.org/10.1145/3452118
14. Gelauff, L., Nikolenko, L., Sakshuwong, S., Fishkin, J., Goel, A., Munagala, K., Siu, A.: Achieving parity with human moderators: a self-moderating platform for online deliberation. In: Boucher, S., Hallin, C.A., Paulson, L. (eds.) The Routledge Handbook of Collective Intelligence for Democracy and Governance, pp. 202–221. Routledge, New York (2023)
15. Rawls, J.: Political Liberalism, Expanded Edition. Columbia University, New York (2005)
16. Ecker, U., Roozenbeek, J., van der Linden, S., Tay, L.Q., Cook, J., Oreskes, N., Lewandowsky, S.: Misinformation poses a bigger threat to democracy than you might think. Nature. **630**, 29–32 (2024). https://doi.org/10.1038/d41586-024-01587-3
17. Coeckelbergh, M.: Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. AI Ethics. **3**(4), 1341–1350 (2023). https://doi.org/10.1007/s43681-022-00239-4
18. Watson, L., Horowitz, M.A., Nieminen, H., Lehtisaari, K., D'Arma, A.: Conclusion: ubiquitous need for epistemic rights and the way forward. In: Horowitz, M.A., Nieminen, H., Lehtisaari, K., D'Arma, A. (eds.) Epistemic Rights in the Era of Digital Disruption, pp. 205–219. Palgrave Macmillan, Cham (2024)
19. Lorenz-Spreen, P., Lewandowsky, S., Sunstein, C.R., Hertwig, R.: How behavioural sciences can promote truth, autonomy and democratic discourse online. Nat. Hum. Behav. **4**, 1102–1109 (2020). https://doi.org/10.1038/s41562-020-0889-7

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.