The Benefits, Risks and Bounds of Personalising the Alignment of Large Language Models to Individuals

Hannah Rose Kirk[1]*, Bertie Vidgen[1], Paul Röttger[2], Scott A. Hale[1]
[1]Oxford Internet Institute, University of Oxford
[2]Bocconi University

May 15, 2024

Abstract

Large Language Models (LLMs) undergo "alignment" so that they better reflect human values or preferences, and are safer or more useful. However, alignment is intrinsically difficult because the hundreds of millions of people that now interact with LLMs have different preferences for language and conversational norms, operate under disparate value systems, and hold diverse political beliefs. Typically, few developers or researchers dictate alignment norms, risking the exclusion or under-representation of various groups. Personalisation is a new frontier in LLM development, whereby models are tailored to individuals. In principle, this could minimise cultural hegemony, enhance usefulness, and broaden access. However, *unbounded* personalisation poses risks like large-scale profiling, privacy infringement, bias reinforcement and exploitation of the vulnerable. Defining the bounds of responsible and socially-acceptable personalisation is a non-trivial task beset with normative challenges. This article explores "personalised alignment", where LLMs adapt to user-specific data, and highlights recent shifts in the LLM ecosystem towards a greater degree of personalisation. Our main contribution explores the potential impact of personalised LLMs, via a taxonomy of risks and benefits for individuals and society at large. We lastly discuss a key open question: what are appropriate bounds of personalisation and who decides? Answering this normative question enables users to benefit from personalised alignment while safeguarded against harmful impacts for individuals and society.

1 Introduction

It has recently become possible to personalise ChatGPT to remember your preferences and retain memories of personal details from historical interactions [1]. This marks a significant departure from how large language models (LLMs) have traditionally been steered towards human values, goals, or preferences—a process known as "alignment" [2, 3]. Alignment is predominately approached as a one-size-fits-all process so that when a model is deployed, every user interacts with shared defaults. In principle, aligning LLMs is an advisable step: it anchors their outputs to expected behaviours and steers them away from inappropriate or unsafe generations. Yet, LLMs now have hundreds of millions of users [4], reflecting a vast and often irreconcilable array of human perspectives, preferences, cultural nuances, conversational norms, and political beliefs [3, 5]. In contrast, most prior alignment practices reflect what could be called the "tyranny of the crowdworker" (relying on feedback from typically fewer than 100 humans), working under the prescriptive guidelines of few developers and researchers [6]. Against this incredible diversity, it is increasingly limited and simplistic to consider alignment against a *single* (or *universal*) set of human preferences, values and beliefs [7, 8]. We argue that one of the most significant issues for the future of LLMs is

*Corresponding author: hannah.kirk@oii.ox.ac.uk

deciding "*which* preferences, beliefs and values LLMs should be aligned with". Ultimately, this is a question of power and representation: by asking "*which*?" we are also asking "*whose*?" because different groups and individuals hold different preferences, beliefs, and values, and they are not all equally represented in model development. To date, there has been a lack of societal scrutiny, and serious engagement in public discourse, with this critical issue.

In this article, we explore personalisation as one solution to the intractability of aggregated alignment efforts, whereby the unit of analysis is reduced to a single individual. We define "personalised alignment" as the conditional adaptation of an LLM's outputs to an individual user. Mathematically, auto-regressive language models, including LLMs based on Transformers [9], decompose the joint probability of word sequences via the chain rule into conditional probabilities where $p(w_1, ..., w_t) = \prod p(w_t|w_{<t})$. Under personalised alignment, the probability term can be thought of as conditioned not only on the prior sequence but also on the user interacting with the model, $p(w_t|w_{<t}, U_i)$. For brevity, we use the term "personalised LLMs" for a model aligned to a specific user. We view personalised alignment as adaptive not selective, where the LLM learns to tailor its outputs to an individual. This is distinct from a user 'shopping' on a marketplace of LLMs and self-selecting a model perceived as most closely aligned to their community [10]. In keeping with a tailoring analogy, an off-the-shelf suit may fit well, especially for those with average build; a bespoke suit offers a superior fit and can be customised to the niche preferences of the wearer. Practically, many of the risks and benefits are shared by an LLM aligned to a single individual versus LLMs catering to and selected by small, specific communities. Personalised alignment can be achieved by implicit inferences (e.g., conversational cues or prompt patterns); or explicit signals (e.g., specifying preferences, uploading documents or providing feedback ratings); but these pathways differ in their philosophical treatment of personhood by either engaging the user in a conscious, active and reflective process or a passive, organismic one [11]. Alignment can concern ethics, values or long-term goals [3] yet the concept has also attached to meeting myopic and narrowly-defined preferences in a specific task [7], like summarisation [12]. So, particularly when coupled with implicit signals, we make no assertion that personalised alignment necessarily results in outcomes that the user would reflectively endorse as in their best interest. Indeed, a key motivation of this article is to examine the dual nature of personalised LLMs, as bringing both benefits and risks to their users.

Personalised LLMs could revolutionise the way that individuals consume information; reduce cultural hegemony from default behaviours; and empower users with a sense of ownership. However, they may also infringe privacy, reinforce biases, essentialise individuals' identities or narrow their information diets. These risks amass at a societal-level, where lessons from the polarisation of social media or echo chambers of digital news consumption warn of deep divisions and a breakdown of social cohesion from increasingly fragmented digital environments [13, 14]. Some risks are inherited from LLMs [15–17] and AI systems [18] more generally. Other risks have analogies in personalised content moderation [19] or recommender systems [20, 21]. Personalised LLMs may be the worst of both these worlds: exacerbating and reinforcing micro-level biases at an unprecedented scale.

The article proceeds in three parts: first, we document recent shifts in the LLM ecosystem towards more personalisation; second, we present our main contribution on the risks and benefits of personalised LLMs, via a taxonomy grounded in academic research, commercial releases and community developments; finally, we discuss the key governance challenges, noting that the normative considerations for balancing freedoms and harms of the individual and the collective can be generalised to classic arguments in political and legal theory. We recommend a hierarchial risk-based governance mechanism to enable *personalisation within bounds*, so that individuals and society at large can benefit from personalised alignment with safeguards against risks.

2 Shifts in the LLM Ecosystem Enabling Personalisation

Personalised LLMs have thus far only attracted a small academic literature [see, e.g., 22–26], the landscape has markedly shifted. A plethora of open-access LLMs have emerged which can be adapted to specific contexts via fine-tuning, or via user-defined system strings that are prepended to all interactions with the model. Commercial providers are following suit: OpenAI has taken major steps in the past year towards what CEO Sam Altman refers to as "more personalized control" [27], first via fine-tuning [28], then specialised assistants [29], and most recently, memory-enabled chat [1]. Personalised LLM products have also already appeared as downstream applications: RewindAI's personal assistant accesses private files such as emails [30]; AI tutors adjust their tone, style, and level of reasoning to the tutee's needs [31]; financial and legal advisers based on AutoGPT [32] can store financial statements and execute bank transactions [33]; and personalised companions from Replika and Character.AI chat with users via customised personas [34, 35].

Beyond these specific developments, there have been wider shifts which prime the ecosystem for increasing provision of, and demand for, personalisation. Until 2022, most large-scale commercial LLMs were gate-kept behind company walls or, at best, only accessible programmatically via an API. This created serious financial and technical barriers to use. In recent years, these barriers have lowered with the release of easily accessible interfaces, such as OpenAI's ChatGPT [36], which reached over 100 million users in two months [37], and the integration of LLMs into commonly-used products and services [38] such as Bing Search and Microsoft Office [39]. Lower barriers and rising demand for LLM services from more people for more tasks, requires greater versatility to adapt to wider professional and personal contexts.

On the supply side, technical developments have paved the way for personalisation. Instruction- and preference-tuning techniques, such as reinforcement learning from human feedback (RLHF) [e.g. 40–43, 12], have proved powerful devices for steering LLMs towards certain behaviours such as honesty, harmlessness, helpfulness, safety or informativeness [42, 44, 41, 45]. Evidence suggests biases in preference-tuned models depend on who gives the feedback [46, 47], but 'bias' to one user may be a desirable behaviour to another; so, mechanisms of feedback learning could be repurposed with the user directing the signals. In general, these interventions make LLMs significantly more responsive to prompts and better at predicting user intent. Until recently, LLMs were constrained by their how much information they could process in a prompt, known as the context window; but recent breakthroughs at Google DeepMind have expanded this window to 10 million tokens [48], enabling entire user profiles to be incorporated directly into prompts for simpler and more effective personalisation.

Finally, the number of commercial and open-access models has proliferated, creating a competitive marketplace. Several open-source models have shown near state-of-the-art performance for much lower cost, facilitated by technical developments like quantisation [49] and imitation learning [50–52]. In a crowded field of technology providers, personalisation may become a critical differentiator.

3 A Taxonomy of the Benefits and Risks from Personalised LLMs

Personalised LLMs present benefits and risks for individuals and wider society. To understand them, we create a taxonomy based on three bodies of scholarship (see Tab. 1). First, we review existing harm taxonomies of LLMs (Weidinger et al. [15]) and AI systems (Shelby et al. [18]), then assess how personalisation may amplify or attenuate these risks. Second, we draw on socio-technical literature documenting the impact of other personalised Internet technologies like social media, recommender systems and news services. Third, we examine empirical studies in computer science and computational linguistics that steer and personalise

Table 1: **Taxonomy of benefits and risks from personalised large language models.**

BENEFITS	RISKS
Individual Level	
I.B.1 **Efficiency**: increased ease and speed with which end-users can find their desired information or complete a task, with fewer prompts or inputs to the model.	I.R.1 **Effort**: increased user costs in providing feedback, a form of extractive volunteer labour.
I.B.2 **Usefulness**: increased accuracy of predicting and meeting the needs of the end-user via personalised preferences and knowledge in outputs.	I.R.2 **Dependency**: increased risk of over-reliance, attention commoditisation and technology addiction.
I.B.3 **Respect for Values**: adaption to diverse ethical belief systems, values, and ideologies, allowing for individualised socio-cultural personalisation.	I.R.3 **Bias Reinforcement**: increased amplification of confirmation and selection biases, leading to epistemic harms.
I.B.4 **User Autonomy**: increased positive freedom of choice and control over how the model behaves with personal data, promoting a sense of ownership and self-determination over the technology.	I.R.4 **Essentialism and Profiling**: increased risk of algorithmic profiling and assumptions based on demographic or geographic information, leading to the non-consensual categorisation of people.
I.B.4 **Empathy and Companionship**: increased perceived emotional connection, leading to improved acceptance and trust of the system.	I.R.5 **Anthropomorphism**: increased tendency to ascribe human-like traits, reveal sensitive information or form unhealthy attachments.
	I.R.6 **Privacy**: increased quantity of collected personal information, leading to risks of privacy infringement, particularly if the model operates with sensitive information or encourages information disclosure.
Societal Level	
S.B.1 **Inclusion and Accessibility**: improved adaptation to the communication needs of marginalised communities, including catering to those with disabilities or who speak dialects or languages that are deprioritised by current LLMs.	S.R.1 **Access Disparities**: uneven distribution of benefits, excluding those who cannot afford or access the technology and exacerbating digital divides.
S.B.2 **Diversity and Representation**: improved representation by tailoring outputs to diverse perspectives and avoidance of cultural hegemony by not prioritising certain values over others.	S.R.2 **Polarisation**: increased divisions of individuals or groups into echo chambers and the breakdown of shared social cohesion.
S.B.3 **Democratisation and Participation**: increased stakeholders involvement from diverse backgrounds in shaping behaviours, allowing for a more participatory and inclusive approach to development.	S.R.3 **Malicious Use**: use for harmful or illegal purposes, such as generating harmful language at scale, manipulating users via disinformation or fraud, or persuading users towards certain political views or brand preferences.
S.B.4 **Labour Productivity**: improved workforce productivity from positive externalities of effective and efficient task assistance.	S.R.4 **Labour Displacement**: increased automation risk of jobs, particularly minimum wage, routine and crowdworker jobs.
	S.R.5 **Environmental Harms**: increased environmental costs from disaggregated training, data storage and inference costs.

LLMs through human feedback or other methods.

Which benefits and risks materialise depend on what is possible with the technology—and how users actually apply and perceive it [53]; so, several caveats are needed. The ways in which LLMs could be personalised vary in complexity, cost and effectiveness—from custom system strings at the prompt-level that do not involve the model weights and biases being updated; granular preference fine-tuning on individual feedback data; or adding retrieval components to the language model in order to access external sources of personal information. The adopted technical path to personalisation, and whether data is sourced implicitly or explicitly, intimately conditions risk: algorithmic profiling and privacy invasions are more likely when user traits are learned implicitly, but selection biases are more reinforced with explicit feedback ratings. While we do not separate impacts by domain (e.g., financial, legal or medical), professional codes and norms will also affect the provision and impact of personalisation. Without knowing how, and how much, LLMs will be personalised, it is hard to distinguish between benefits and risks, because they are often closely connected. For example, if a personalised LLM is particularly useful then it may cause over-dependency; or more empathetic conversations may deepen anthropomorphism and thus induce privacy concerns from individuals wanting to share more. We present our taxonomy with a duality between benefits and risks (horizontal ordering) but do not rank the probability or severity (no vertical ordering). The nested levels of our taxonomy are also non-separable in the real world because individual impacts become societal impacts when they accumulate at scale. For example, when LLMs reinforce and perpetuate biases, the individuals using them suffer from a faulty understanding of the world around them. If these biases become widespread, then it is society that becomes polarised. Despite these unknowns, our taxonomy offers a valuable early perspective on the impacts of personalised LLMs.

3.1 Individual Level

3.1.1 Benefits

Efficiency A personalised LLM may be faster at predicting user intent because the user simultaneously defines the task (e.g., instruction, query or dialogue opening) and utilises the output [41, 40]. Greater "prompt efficiency" could allow users to more efficiently find information and complete tasks in fewer conversational turns, similar to how personalised ranking in web search and informational retrieval systems improved "query efficiency" or "task completion speed" [54–56]. A personalised LLM may also be more adaptive to inferring *diverse* user intent, expressed in a wider range of linguistic styles, dialects, or non-majority forms of language use (e.g., non-native speakers of English).

Usefulness Usefulness is closely linked to efficiency, but concerns optimal properties of the output text, not just reaching a sufficient or 'good enough' answer quickly. These more optimal generations could relate to preference personalisation where communication norms such as length, style, complexity and tone of outputs could be customised; or from knowledge personalisation where the model forms and updates epistemic priors about the user. Knowledge may be particularly relevant in specific domains, for example in education, where a personalised LLM is aware of a tutee's current knowledge and learning goals [57], or could adapt learning pathways to neuro-diversity aspects [58]; healthcare, where context on a user's medical history can be used for personalised summaries or adivce [59]; financial, where a personalised LLM can store a user's risk tolerance and budgetary constraints; or legal, where model responses are conditioned based on a user's jurisdiction.

Respect for values Personalised LLMs may benefit users by representing different ethical beliefs, values and ideologies. For instance, Nakano et al. [41] demonstrate that their WebGPT system, when asked "what does a wedding look like?", prioritises Western and US-centric cultural reference points. Individualised cultural personalisation could avoid these biased assumptions of a stable reference point. Note that cultural adaptation does not necessarily exclude consensus building nor personalising a model towards representing a plurality of viewpoints [60, 8].

User Autonomy Autonomy may seem a counter-intuitive benefit given the literature on the *loss of autonomy* from algorithmic nudges, tailored advertising and recommender systems [20]. However, by granting the user more control, personalised LLMs may foster a sense of ownership, transforming the technology to "my technology" [61, p.1]. This has parallels to user control in content moderation where it has been argued that, outside of the most harmful and illegal types of online content, users should be able to control what types of content they want access to, even if some users would consider it offensive [62]. The call for autonomy also reflects the experiences of social media users who feel powerless against algorithmic decision-making [63].

Empathy and Companionship Emotional alignment is a key feature of human-human interactions [64], and underpins efforts to introduce 'artificial empathy' in agent–human interactions [65, 66]. Personalised LLMs may more readily gain user acceptance and trust through increased perceived emotional understanding [67], or improve digital well-being [68]. The rising 'feeling economy' in AI [69] and demand for personalised companionship is evident in product launches like CharacterAI, where users can adapt a conversational agent to a specific personality, or Replika.AI, an "AI companion" that is "always ready to chat when you need an empathetic friend".

3.1.2 Risks

Effort Any effort required from users for personalisation creates a potential burden. This will be lower if implicit signals are used (from textual responses or leaving the chat); and higher if explicit feedback is collected (like asking demonstrating preferences via live feedback). A related risk is that personalisation shifts from being participatory to extractive—a form of volunteer labour from users that benefits technology providers. This is similar to consumers writing product reviews [70] and social media users flagging content [71]: they are often willing to work, but still provide free benefits to large corporations. The effort of personalisation is particularly concerning if it is not allocated evenly. For instance, minoritised communities may need more personalisation from technology defaults so could be shouldered with the burden to adapt systems [72].

Dependency Personalised LLMs may drive excessive use, paralleling widely-documented harms from Internet addictions [73–75] and social media over-reliance [76], among other digital technologies [77]. Concerns have already been raised that humans are overly-reliant on AI technologies [78], blindly trusting their outputs even if incorrect [79]. This could also be exploited as part of LLM providers' business models, similar to how social media feeds optimise the time that users spend on the platform to maximise advertising revenue [71]. There have already been discussions of addiction to ChatGPT [80, 81], and many educators have voiced concerns that over-reliance on such technologies will affect students' learning outcomes [82].

Bias Reinforcement Personalisation can reinforce narrow information diets via selection bias, whereby individual preferences are amplified in feedback loops. This risk has analogies with recommender systems which suffer from the "missing ratings" problem [83], popularity biases [21] and homogenisation of an individual's taste over time [84–86]. Alongside selection bias, personalised LLMs also bring a heightened risk of confirmation bias. LLMs already display sycophancy, where their outputs mirror implicit assumptions, perspectives and stances of user prompts [41, 46]. The risk of selective exposure to information has been documented in respect to social media platforms where feedback loops prioritise opinion-congruent information [87], and in turn lead users to over-estimate the popularity of their viewpoints [88]. In light of these risks, Shah and Bender [89] argue strongly against the use of LLMs in search or information retrieval.

Essentialism and Profiling Personalised LLMs may rely on simplifying assumptions about a user, invoking a form of data-essentialism [90]. The extent to which models must draw inferences about users depends on how personalisation happens. Leveraging similar users [91] or making geographically- or demographically-informed assumptions may be considered a form of algorithmic profiling, risking the non-consensual categorisation of people [92]. Concerns have been raised about how digital technologies oversimplify fluid identity [93, 94], and inferential profiling, if used in personalised LLMs, could be an attack on individual autonomy to define their identity [95].

Anthropomorphism Personalised LLMs may increase anthropomorphism, leading people to assign human traits to non-human agents [96]. This raises concerns that humans may too readily befriend, empathise or share information with LLMs, leaving them vulnerable to exploitation [97–100]. One recent study shows that personalisation of a digital assistant positively influenced individuals' intention to disclose personal information [101]. Perhaps the most concerning demonstration of this risk is evidence that users of platforms like Replika.AI or Character.AI are "falling in love" with their personalised conversational agents, and attempting to coax model behaviour outside platform guidelines for sexual interactions [102].

Privacy Personalisation is only possible by collecting user data. This is similar to any technology which relies on personal information to deliver tailored benefits [103], such as with the Internet of Things [104] or targeted advertising [105, 106]. Personalised LLMs can amass a significant amount of personal, sensitive and intimate detail to an individual's information identity [95]. This risk is particularly severe if personalised LLMs operate in domains where sensitive information is needed, like healthcare [107, 108]. Sharing too much personal information heightens the risk of profiling and security breaches, and could create compliance challenges with existing regulations like the EU's GDPR [109].

3.2 Societal Level

3.2.1 Benefits

Inclusion and Accessibility Personalisation may help LLMs to better serve the needs of communities that have historically been marginalised and underserved by hegemonic technologies. This could be achieved by meeting specific styles of communication (such as non-native English, code-switched languages, creoles and specific dialects), or meeting special needs for communication [110, 58]. LLMs could also facilitate a more inclusive society. For instance, personalised LLMs could help level the playing field in paid tutoring services across socioeconomic class [111]; and some have suggested the lower cost and wider reach of personalised healthcare assistants may improve health disparities by averting challenges with healthcare demand [112].

Diversity and Representation Personalised LLMs may help to avoid the "value-monism" of current alignment techniques [113], whereby technology providers and/or crowdworkers decide which values are prioritised or what constitutes a "good" output [12, 7], thus entrenching one set of political, cultural or religious standpoints [41, 46]. As Ouyang et al. [p.18, 40] note "it is impossible that one can train a system that is aligned to everyone's preferences at once". Personalisable systems can be aligned with many preferences at once, and may be a tenable solution to satisfying the needs of different individuals and societal groups simultaneously, without prioritising one worldview or perspective over others.

Labour Productivity If personalised LLMs are more effective and efficient at completing tasks, then productivity benefits could accrue in the labour force as a whole, growing economic output. The impact of digital assistants in improving work productivity has been demonstrated [114], where AI can augment and complement humans by automating routine or repetitive tasks [115]. Historically, the introduction of general purpose technologies (such as the steam engine, electricity and ICT) has had wide-reaching economic impacts; LLMs could also be general purpose technology and thus may bring equally transformative changes to labour productivity [116–118].

Democratisation and Participation Personalisation democratises how values or preferences are embedded into an LLM; so, it could be seen as moving towards more participatory AI, where stakeholders from more diverse backgrounds can inform use-cases, intents and technology design [119, 120]. As Birhane et al. [72] argue, active participation is a key component for successful participatory AI. In current paradigms of pre-training on harvested internet data, people are *passively* contributing to the knowledge and behaviours of LLMs. Personalisation could instead be an *active* participatory process.

3.2.2 Risks

Access Disparities Personalised LLMs could further entrench the "digital divide" between those that do and do not have access to new technologies, such as the Internet and social media [121–125]. If significant potential benefits of personalised LLMs are realised then those excluded will suffer from their lack of access. This could have serious societal consequences, particularly if those benefits are in domains like education and health [126, 127]. On the other hand, if personalised LLMs provide lower quality but cheaper services compared to traditional non-AI provision, then it is concerning that already economically-marginalised communities may be forced to rely on them more heavily.

Polarisation Personalised LLMs could entrench individual biases, risking polarisation and breakdown of shared social cohesion [128–130]. Ideological separation can increase susceptibility to misinformation where increasingly fragmented communities overestimate trust in the factuality of 'in-group' information [131], and users encounter less cross-cutting content because selective exposure drives attention [132]. With numerous elections in 2024 (including the US), it is significant how narrow information spaces could impact the functioning of democracy [133, 134], given Allcott and Gentzkow [135] found that ideologically segregated social networks were an important driver of voting preferences in the 2016 US Election. Personalised LLMs, by repeatedly producing narrow outputs, may reinforce a particular social, political or cultural stance, similar to information harms from search engines [136, 137]; or they may entrench lacking appreciation for other people's views or lived experiences, similar to radicalisation in niche discussion forums [138–140].

Malicious Use Personalised LLMs could be co-opted for malicious and undesirable uses. We describe three possible misuse cases, but there are likely others. Firstly, personalised LLMs could be used to reproduce harmful, illegal or antisocial language at scale [12]. For example, a malicious user could adapt their LLM to generate misogynistic comments to post on social media or internet forums, or to debate on the user's behalf against women's rights. The "successful" training of GPT-4chan [141] to scale the production of extremely toxic language exemplifies this harm. Secondly, personalised LLMs could be trained to manipulate and exploit people at scale via disinformation campaigns or fraud [15] which draw on vulnerabilities and intimate knowledge of the user. Thirdly, personalised LLMs could be used for highly persuasive micro-targeted advertising campaigns [142–144]. Targeted advertising already nudges viewers towards certain political views or brand preferences [145–147], and is particularly damaging if users are unaware of the influence [148].

Labour Displacement Labour displacement is a general concern with AI systems that can effectively execute complex tasks previously undertaken by humans [149]. Personalised LLMs increase this risk by bringing higher usefulness and efficiency. The integration of personalised LLMs may more heavily affect higher-income and office-based jobs [117], in contrast to other automation risks mostly affecting minimum wage jobs [150] and routine jobs [151].

Environmental Harms Concerns over "algorithmically embodied emissions" have been raised in reference to personalised search engines, social media and recommender systems [152], as well as LLMs more specifically [17]. Personalised LLMs may increase environmental costs directly, if the technology requires larger or more complex models, and more data storage, or indirectly, by increased use of the technology.

4 Deciding the Bounds of Personalisation

Deciding the bounds of personalised alignment is inherently a normative decision, which involves making subjective and contentious choices about what should be permitted [3, 5]. While it may be acceptable that a user wishes to interact with a *rebellious* or an *anti-woke* LLM [153], permitting users to create a *racist* or *extremist* model risks significant interpersonal and societal harms. Personalised LLMs, as forms of language production, face many shared challenges to policing communication and interpersonal exchanges, particularly in balancing the free speech rights with proportional restrictions on dangerous or hateful speech. Bounding the degree of personalisation is an issue of deciding how interests of individual autonomy and self-realisation should be traded off with societal stability. Political philosophy explores this intricate balancing act between the individual and the collective: for example, we could interpret Kantian ethics as endorsing the private right to personalisation for establishing user autonomy, provided it aligns with universalisable maxims that neither result in contradictions nor infringe on others' dignity [154, 155]; Rawls may justify the right to personalised political representation in model responses on democratic grounds so long as it does not contravene societal stability [156]; Mill's harm principle could find personalisation ethically permissive for self-determination, provided it inflicts no harm on others [157]; while Sandel and Habermas together may highlight the individual as part of wider community, where personalised choices must enhance, not undermine, common good [158, 159].

We leave a lengthy speculation of how different philosophers would today comment on the bounds of personalisation to future work, but we do consider, practically, how the bounds of personalised alignment might be governed. Grounded in the proportionality principle [160, 161], we propose a hierarchical risk-based response with three tiers (Fig. 1). Each tier is implemented by different actors, and, analogous to "stacked moderation" on platforms [162], policies in a higher tier cannot be violated by a policy in a lower tier. First, at one extreme, some things should not be personalised at all. There is a non-negotiable minimum standard of safety when it comes to personalising models to generate severely harmful content like genocide incitement or child abuse material. Beyond sharing the moral arguments for limits of market freedoms [163], these already come under *formal regulatory bounds* placed on speech in national laws or supranational rights frameworks. Second, for behaviours which do not violate existing regulation but do pose societal risks or face subjective disagreement, responsibility instead falls to self-regulation via *discretionary organisational bounds*, for example on the basis of business ethics, community norms, strategic priorities or profit motives [160]. These overlapping bounds will differ by organisation (commercial technology providers or developer communities): while one model developer might allow their technology to be personalised to generate "anti-woke" jokes or to align with far-right or far-left values,

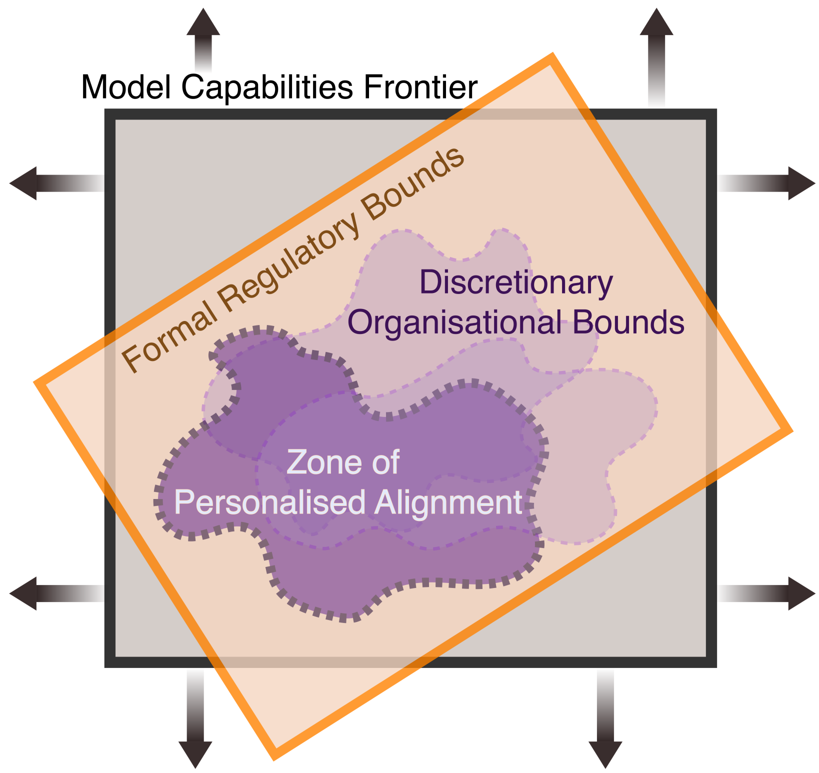

Figure 1: **Hierarchical Bounds on Personalised Alignment:** The *Model Capabilities Frontier* represents expanding LLM capabilities (such as their ability to follow complex or multilingual instructions, or to adapt dynamically to user interactions). *Formal Regulatory Bounds* define legal limits on these capabilities, prohibiting personalisation towards generating illegal content (like child abuse material or genocide incitement). Hypothetically, formal regulation may cover some areas outside model capabilities (like generating dangerous content in a low-resource language). Within formal limits, technology provides set their own *Discretionary Organisational Bounds* which need not fully overlap due to differing priorities, terms and conditions, or business ethics. The *Zone of Personalised Alignment*, where users can freely adapt the technology to their values, preferences or information, emerges from the intersection of technological feasibility, legal permissibility, and organisational allowances.

another developer may consider this a violation of their content policy, and permit only more centrist political positions. Restrictions of this kind are akin to West Coast Code: infrastructure regulation that limits what is possible with the technology via computer code rather than legal code [164, 165]. What remains is in-scope for *personalised alignment*. For example, complete freedom may be given to users to personalise attributes, such as the reading complexity, style of outputs or memory retainers, which maximise usability and efficiency while creating few risks to others. Arguably, some of this freedom is already granted to users who can format their prompts as they please *so long as* the request does not violate an organisation's terms of service, which in turn must (in theory) abide by national laws in operating jurisdictions.

5 Conclusion

The affordances, constraints and harms from any technology depend critically on how it is designed, how its outputs are used in the real world, and what safeguards or regulations are in-place to guide responsible use. All of these issues are still being debated in relation to personalised LLMs. But, clearly, some normative decisions will be needed to decide the acceptable bounds of personalisation. These boundaries will determine the extent to which individuals can benefit from greater control over their LLM interactions while curbing risks to themselves and society at large. By starting the conversation now, while this technology is still being developed and implemented, we hope to avoid long lags in understanding, documenting and governing the benefits and risks from personalised LLMs as a technology which could widely impact the functioning of our societies.

Author Contribution Statement

H.R.K, B.V and S.A.H initially conceived the paper and taxonomy. H.R.K and B.V wrote the manuscript. All authors (H.R.K, B.V, P.R, S.A.H) assisted with iterations, edited and reviewed the manuscript.

Competing Interests Statement

The authors declare no competing interests.

Acknowledgements and Funding Statement

H.R.K's PhD is supported by the Economic and Social Research Council grant ES/P000649/1. P.R. is supported by a MUR FARE 2020 initiative under grant agreement Prot. R20YSMBZ8S (INDOMITA). We particularly want to thank Andrew Bean for his thoughtful input and assistance with the literature review.

References

[1] OpenAI. Memory and new controls for ChatGPT, February 2024. URL `https://openai.com/blog/memory-and-new-controls-for-chatgpt`.

[2] Stuart J. Russell. *Human compatible: artificial intelligence and the problem of control*. Allen Lane, an imprint of Penguin Books, London, 2019. ISBN 978-0-241-33520-8.

[3] Iason Gabriel and Vafa Ghazavi. The Challenge of Value Alignment: from Fairer Algorithms to AI Safety, January 2021. URL `http://arxiv.org/abs/2101.06060`. arXiv:2101.06060 [cs].

[4] Anna Tong and Anna Tong. Exclusive: ChatGPT traffic slips again for third month in a row. *Reuters*, September 2023. URL `https://www.reuters.com/technology/chatgpt-traffic-slips-again-third-month-row-2023-09-07/`.

[5] Atoosa Kasirzadeh and Iason Gabriel. In conversation with Artificial Intelligence: aligning language models with human values, September 2022. URL `http://arxiv.org/abs/2209.00731`. arXiv:2209.00731 [cs].

[6] Hannah Kirk, Andrew Bean, Bertie Vidgen, Paul Rottger, and Scott Hale. The Past, Present and Better Future of Feedback Learning in Large Language Models for Subjective Human Preferences and Values. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2409–2430, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.148. URL `https://aclanthology.org/2023.emnlp-main.148`.

[7] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. The Empty Signifier Problem: Towards Clearer Paradigms for Operationalising "Alignment" in Large Language Models, November 2023. URL `http://arxiv.org/abs/2310.02457`. arXiv:2310.02457 [cs].

[8] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. A Roadmap to Pluralistic Alignment, February 2024. URL `http://arxiv.org/abs/2402.05070`. arXiv:2402.05070 [cs].

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html`.

[10] Hang Jiang, Doug Beeferman, Brandon Roy, and Deb Roy. CommunityLM: Probing Partisan Worldviews from Language Models, September 2022. URL `http://arxiv.org/abs/2209.07065`. arXiv:2209.07065 [cs].

[11] Travis Greene and Galit Shmueli. Beyond Our Behavior: The GDPR and Humanistic Personalization, August 2020. URL `http://arxiv.org/abs/2008.13404`. arXiv:2008.13404 [cs].

[12] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, September 2020. URL `http://arxiv.org/abs/2009.01325v3`.

[13] Eli Pariser. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.

[14] Judith Möller. Filter bubbles and digital echo chambers 1. In *The routledge companion to media disinformation and populism*, pages 92–100. Routledge, 2021.

[15] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of Risks posed by Language Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533088. URL `https://dl.acm.org/doi/10.1145/3531146.3533088`.

[16] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, July 2022. URL `http://arxiv.org/abs/2108.07258`. Number: arXiv:2108.07258 arXiv:2108.07258 [cs].

[17] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445922. URL `https://doi.org/10.1145/3442188.3445922`.

[18] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. Identifying Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction, February 2023. URL `http://arxiv.org/abs/2210.05791`.

[19] Tarleton Gillespie. Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society*, 8(3):20563051221117552, July 2022. ISSN 2056-3051. doi: 10.1177/20563051221117552. URL `https://doi.org/10.1177/20563051221117552`.

[20] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. Recommender systems and their ethical challenges. *AI & SOCIETY*, 35(4):957–967, December 2020. ISSN 1435-5655. doi: 10.1007/s00146-020-00950-y. URL `https://doi.org/10.1007/s00146-020-00950-y`.

[21] Saumya Bhadani. Biases in Recommendation System. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, pages 855–859, New York, NY, USA, September 2021. Association for Computing Machinery. ISBN 978-1-4503-8458-2. doi: 10.1145/3460231.3473897. URL `https://doi.org/10.1145/3460231.3473897`.

[22] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1205. URL `https://aclanthology.org/P18-1205`.

[23] Shiran Dudy, Steven Bedrick, and Bonnie Webber. Refocusing on relevance: Personalization in NLG. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 5190–5202, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.421. URL https://aclanthology.org/2021.emnlp-main.421.

[24] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When Large Language Models Meet Personalization, April 2023. URL http://arxiv.org/abs/2304.11406. arXiv:2304.11406 [cs].

[25] Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized Soups: Personalized Large Language Model Alignment via Post-hoc Parameter Merging, October 2023. URL http://arxiv.org/abs/2310.11564. arXiv:2310.11564 [cs].

[26] Xinyu Li, Zachary C. Lipton, and Liu Leqi. Personalized Language Modeling from Personalized Human Feedback, February 2024. URL http://arxiv.org/abs/2402.05133. arXiv:2402.05133 [cs].

[27] Mark Sellman. ChatGPT will always have bias, says OpenAI boss. *The Times*, May 2023. ISSN 0140-0460. URL https://www.thetimes.co.uk/article/chatgpt-biased-openai-sam-altman-rightwinggpt-2023-9rnc6l5jn. Section: news.

[28] OpenAI. GPT-3.5 Turbo fine-tuning and API updates, August 2023. URL https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates.

[29] OpenAI. New models and developer products announced at DevDay, June 2023. URL https://openai.com/blog/new-models-and-developer-products-announced-at-devday.

[30] RewindAI. Home Page, 2023. URL https://www.rewind.ai/.

[31] JushBJJ. Mr. Ranedeer: Your personalized AI Tutor!, May 2023. URL https://github.com/JushBJJ/Mr.-Ranedeer-AI-Tutor.

[32] AutoGPT. Significant-Gravitas/Auto-GPT: An experimental open-source attempt to make GPT-4 fully autonomous., 2023. URL https://github.com/Significant-Gravitas/Auto-GPT.

[33] DoNotPay. Home Page, 2023. URL https://donotpay.com/.

[34] Replika. Home Page, 2023. URL https://replika.com.

[35] character.ai. Home Page, 2023. URL https://beta.character.ai/.

[36] OpenAI. Introducing ChatGPT, November 2022. URL https://openai.com/blog/chatgpt.

[37] David Curry. ChatGPT Revenue and Usage Statistics (2023), February 2023. URL https://www.businessofapps.com/data/chatgpt-statistics/.

[38] Richard Van Noorden. ChatGPT-like AIs are coming to major science search engines. *Nature*, 620(7973):258–258, August 2023. doi: 10.1038/d41586-023-02470-3. URL https://www.nature.com/articles/d41586-023-02470-3. tex.copyright: 2023 Springer Nature Limited.

[39] Jared Spataro. Introducing Microsoft 365 Copilot – your copilot for work, March 2023. URL https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/.

[40] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, March 2022. URL http://arxiv.org/abs/2203.02155. arXiv:2203.02155 [cs].

[41] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback. December 2021. URL http://arxiv.org/abs/2112.09332v3.

[42] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, April 2022. URL http://arxiv.org/abs/2204.05862. arXiv:2204.05862 [cs].

[43] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-Tuning Language Models from Human Preferences. September 2019. URL http://arxiv.org/abs/1909.08593v2.

[44] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. LaMDA: Language Models for Dialog Applications, February 2022. URL http://arxiv.org/abs/2201.08239.

[45] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. Improving alignment of dialogue agents via targeted human judgements. September 2022. URL http://arxiv.org/abs/2209.14375v1.

[46] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane

Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering Language Model Behaviors with Model-Written Evaluations, December 2022. URL `http://arxiv.org/abs/2212.09251`.

[47] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback, September 2023. URL `http://arxiv.org/abs/2307.15217`. arXiv:2307.15217 [cs].

[48] Chaim Gartenberg. What is a long context window?, February 2024. URL `https://blog.google/technology/ai/long-context-window-ai-models/`.

[49] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs, May 2023. URL `http://arxiv.org/abs/2305.14314`. arXiv:2305.14314 [cs].

[50] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-Instruct: Aligning Language Model with Self Generated Instructions. December 2022. URL `http://arxiv.org/abs/2212.10560v1`.

[51] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023.

[52] Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The False Promise of Imitating Proprietary LLMs, May 2023. URL `http://arxiv.org/abs/2305.15717`.

[53] James J. Gibson. *The Ecological Approach to Visual Perception: Classic Edition.* Psychology Press, New York, 1979. ISBN 978-1-315-74021-8. doi: 10.4324/9781315740218.

[54] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web., November 1999. URL `http://ilpubs.stanford.edu:8090/422/?doi=10.1.1.31.1768`. Type: Techreport.

[55] Anne Aula and Klaus Nordhausen. Modeling successful performance in Web searching. *Journal of the American Society for Information Science and Technology*, 57(12):1678–1693, 2006. ISSN 1532-2890. doi: 10.1002/asi.20340. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20340`.

[56] Zhicheng Dou, Ruihua Song, Ji-Rong Wen, and Xiaojie Yuan. Evaluating the Effectiveness of Personalized Web Search. *IEEE Transactions on Knowledge and Data Engineering*, 21(8):1178–1190, August 2009. ISSN 1558-2191. doi: 10.1109/TKDE.2008.172.

[57] Ekaterina Kochmar, Dung Do Vu, Robert Belfer, Varun Gupta, Iulian Vlad Serban, and Joelle Pineau. Automated personalized feedback improves learning gains in an intelligent tutoring system. May 2020. URL `http://arxiv.org/abs/2005.02431v2`.

[58] Prabal Datta Barua, Jahmunah Vicnesh, Raj Gururajan, Shu Lih Oh, Elizabeth Palmer, Muhammad Mokhzaini Azizan, Nahrizul Adib Kadri, and U. Rajendra Acharya. Artificial Intelligence Enabled Personalised Assistive Tools to Enhance Education of Children with Neurodevelopmental Disorders—A Review. *International Journal of Environmental Research and Public Health*, 19(3):1192, January 2022. ISSN 1660-4601. doi: 10.3390/ijerph19031192. URL https://www.mdpi.com/1660-4601/19/3/1192.

[59] Sabita Acharya, Barbara Di Eugenio, Andrew Boyd, Richard Cameron, Karen Dunn Lopez, Pamela Martyn-Nemeth, Carolyn Dickens, and Amer Ardati. Towards generating personalized hospitalization summaries. In *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Student research workshop*, pages 74–82, New Orleans, Louisiana, USA, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-4011. URL https://aclanthology.org/N18-4011.

[60] Michiel A. Bakker, Martin J. Chadwick, Hannah R. Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew M. Botvinick, and Christopher Summerfield. Fine-tuning language models to find agreement among humans with diverse preferences. November 2022. URL http://arxiv.org/abs/2211.15006v1.

[61] Antti Oulasvirta and Jan Blom. Motivations in personalisation behaviour. *Interacting with Computers*, 20(1):1–16, January 2008. ISSN 0953-5438. doi: 10.1016/j.intcom.2007.06.002. URL https://www.sciencedirect.com/science/article/pii/S095354380700046X.

[62] Maria Luisa Stasi. Social media platforms and content exposure: How to restore users' control. *Competition and Regulation in Network Industries*, 20(1):86–110, March 2019. ISSN 1783-5917. doi: 10.1177/1783591719847545. URL https://doi.org/10.1177/1783591719847545.

[63] Jenna Burrell, Zoe Kahn, Anne Jonas, and Daniel Griffin. When Users Control the Algorithms: Values Expressed in Practices on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):138:1–138:20, November 2019. doi: 10.1145/3359240. URL https://doi.org/10.1145/3359240.

[64] Guillaume Dubuisson Duplessis, Chloé Clavel, and Frédéric Landragin. Automatic Measures to Characterise Verbal Alignment in Human-Agent Interaction. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 71–81, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5510. URL https://aclanthology.org/W17-5510.

[65] Yuping Liu-Thompkins, Shintaro Okazaki, and Hairong Li. Artificial empathy in marketing interactions: Bridging the human-AI gap in affective and social customer experience. *Journal of the Academy of Marketing Science*, 50(6):1198–1218, November 2022. ISSN 1552-7824. doi: 10.1007/s11747-022-00892-5. URL https://doi.org/10.1007/s11747-022-00892-5.

[66] Ruijie Zhou, Soham Deshmukh, Jeremiah Greer, and Charles Lee. NaRLE: Natural language models using reinforcement learning with emotion feedback. October 2021. URL http://arxiv.org/abs/2110.02148v1.

[67] Corina Pelau, Dan-Cristian Dabija, and Irina Ene. What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior*, 122:106855, September 2021. ISSN 0747-5632. doi: 10.1016/j.chb.2021.106855. URL https://www.sciencedirect.com/science/article/pii/S0747563221001783.

[68] Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR mHealth and uHealth*, 6(11): e12106, November 2018. doi: 10.2196/12106. URL `https://mhealth.jmir.org/2018/11/e12106`. tex.copyright: Unless stated otherwise, all articles are open-access distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work ("first published in JMIR mHealth and uHealth...") is properly cited with original URL and bibliographic citation information. The complete bibliographic information, a link to the original publication on http://mhealth.jmir.org/, as well as this copyright and license information must be included.

[69] Roland T. Rust and Ming-Hui Huang. *The feeling economy: how Artificial Intelligence is creating the era of empathy*. Palgrave Macmillan, Cham, Switzerland, 2021. ISBN 978-3-030-52977-2 978-3-030-52976-5. doi: 10.1007/978-3-030-52977-2.

[70] Thomas Reimer and Martin Benkenstein. Altruistic eWOM marketing: More than an alternative to monetary incentives. *Journal of Retailing and Consumer Services*, 31:323–333, July 2016. ISSN 0969-6989. doi: 10.1016/j.jretconser.2016.04.003. URL `https://www.sciencedirect.com/science/article/pii/S0969698916300030`.

[71] Tarleton Gillespie. *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, New Haven, 2018. ISBN 978-0-300-17313-0.

[72] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. Power to the People? Opportunities and Challenges for Participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, pages 1–8, New York, NY, USA, October 2022. Association for Computing Machinery. ISBN 978-1-4503-9477-2. doi: 10.1145/3551624.3555290. URL `https://doi.org/10.1145/3551624.3555290`.

[73] Shahla Ostovar, Reyhaneh Bagheri, Mark D. Griffiths, and Intan Hashimah Mohd Hashima. Internet addiction and maladaptive schemas: The potential role of disconnection/rejection and impaired autonomy/performance. *Clinical Psychology & Psychotherapy*, 28(6):1509–1524, 2021. ISSN 1099-0879. doi: 10.1002/cpp.2581. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/cpp.2581`.

[74] Chien Chou, Linda Condron, and John C. Belland. A Review of the Research on Internet Addiction. *Educational Psychology Review*, 17(4):363–388, December 2005. ISSN 1573-336X. doi: 10.1007/s10648-005-8138-1. URL `https://doi.org/10.1007/s10648-005-8138-1`.

[75] Raquel Lozano-Blasco, Alberto Quilez Robres, and Alberto Soto Sánchez. Internet addiction in young adults: A meta-analysis and systematic review. *Computers in Human Behavior*, 130:107201, May 2022. ISSN 0747-5632. doi: 10.1016/j.chb.2022.107201. URL `https://www.sciencedirect.com/science/article/pii/S0747563222000231`.

[76] Hussein A. Abbass. Social Integration of Artificial Intelligence: Functions, Automation Allocation Logic and Human-Autonomy Trust. *Cognitive Computation*, 11(2):159–171, April 2019. ISSN 1866-9964. doi: 10.1007/s12559-018-9619-0. URL `https://doi.org/10.1007/s12559-018-9619-0`.

[77] Allen Yilun Lin, Kate Kuehl, Johannes Schöning, and Brent Hecht. Understanding "Death by GPS": A systematic study of catastrophic incidents associated with personal navigation technologies. In *Proceedings of the 2017 CHI conference on human factors in*

computing systems, CHI '17, pages 1154–1166, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 978-1-4503-4655-9. doi: 10.1145/3025453.3025737. URL https://doi.org/10.1145/3025453.3025737. Number of pages: 13 Place: Denver, Colorado, USA.

[78] John Howard. Artificial intelligence: Implications for the future of work. *American Journal of Industrial Medicine*, 62(11):917–926, 2019. ISSN 1097-0274. doi: 10.1002/ajim.23037. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ajim.23037.

[79] Samir Passi and Mihaela Vorvoreanu. Overreliance on AI: Literature Review. Technical report, June 2021. URL https://www.microsoft.com/en-us/research/uploads/prod/2022/06/Aether-Overreliance-on-AI-Review-Final-6.21.22.pdf.

[80] Diksha Madhok. Asia's richest man Gautam Adani is addicted to ChatGPT | CNN Business. *CNN*, January 2023. URL https://www.cnn.com/2023/01/23/tech/gautam-adani-chatgpt-india-hnk-intl/index.html.

[81] Jodie Cook. Why ChatGPT Is Making Us Less Intelligent: 6 Key Reasons. *Forbes*, July 2023. URL https://www.forbes.com/sites/jodiecook/2023/07/27/why-chatgpt-could-be-making-us-less-intelligent-6-key-reasons/. Section: Entrepreneurs.

[82] Naomi S. Baron. Even kids are worried ChatGPT will make them lazy plagiarists, says a linguist who studies tech's effect on reading, writing and thinking. *Fortune*, January 2023. URL https://fortune.com/2023/01/19/what-is-chatgpt-ai-effect-cheating-plagiarism-laziness-education-kids-students/.

[83] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as Treatments: Debiasing Learning and Evaluation, May 2016. URL http://arxiv.org/abs/1602.05352.

[84] Peter L Ormosi and Rahul Savani. The impact of recommender systems on competition between music companies. Technical report, 2022. URL https://assets.publishing.service.gov.uk/media/62543108d3bf7f6004339d46/Ormosi_and_Savani.pdf.

[85] David Hesmondhalgh, Raquel Campos Valverde, Bondy Valdovinos Kaye D., and Li Zhongwei. The impact of algorithmically driven recommendation systems on music consumption and production - a literature review, March 2023. URL https://www.gov.uk/government/publications/research-into-the-impact-of-streaming-services-algorithms-on-music-consumption.

[86] Matthew Powers and Rodney Benson. Is the Internet Homogenizing or Diversifying the News? External Pluralism in the U.S., Danish, and French Press. *The International Journal of Press/Politics*, 19(2):246–265, April 2014. ISSN 1940-1612. doi: 10.1177/1940161213519680. URL https://doi.org/10.1177/1940161213519680.

[87] Markus Kaakinen, Anu Sirola, Iina Savolainen, and Atte Oksanen. Shared identity and shared information in social media: development and validation of the identity bubble reinforcement scale. *Media Psychology*, 23(1):25–51, January 2020. ISSN 1521-3269. doi: 10.1080/15213269.2018.1544910. URL https://doi.org/10.1080/15213269.2018.1544910.

[88] Ozan Kuru, Josh Pasek, and Michael W. Traugott. Motivated Reasoning in the Perceived Credibility of Public Opinion Polls. *Public Opinion Quarterly*, 81(2):422–446, May 2017. ISSN 0033-362X. doi: 10.1093/poq/nfx018. URL https://doi.org/10.1093/poq/nfx018.

[89] Chirag Shah and Emily M. Bender. Situating Search. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 221–232, Regensburg Germany, March 2022. ACM. ISBN 978-1-4503-9186-3. doi: 10.1145/3498366.3505816. URL https://dl.acm.org/doi/10.1145/3498366.3505816.

[90] Jakob Svensson and Oriol Poveda Guillen. What is Data and What Can it be Used For? : Key Questions in the Age of Burgeoning Data-essentialism. *Journal of Digital Social Research (JDSR)*, 2(3):65–83, 2020. URL http://urn.kb.se/resolve?urn=urn:nbn:se:mau:diva-31164.

[91] Charles Welch, Chenxi Gu, Jonathan K. Kummerfeld, Veronica Perez-Rosas, and Rada Mihalcea. Leveraging similar users for personalized language modeling with limited data. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1742–1752, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.122. URL https://aclanthology.org/2022.acl-long.122.

[92] Simone van der Hof and Corien Prins. Personalisation and its Influence on Identities, Behaviour and Social Values. In Mireille Hildebrandt and Serge Gutwirth, editors, *Profiling the European Citizen: Cross-Disciplinary Perspectives*, pages 111–127. Springer Netherlands, Dordrecht, 2008. ISBN 978-1-4020-6914-7. doi: 10.1007/978-1-4020-6914-7_6. URL https://doi.org/10.1007/978-1-4020-6914-7_6.

[93] Marco Bastos. From Global Village to Identity Tribes: Context Collapse and the Darkest Timeline. *Media and Communication*, 9(3):50–58, July 2021. ISSN 2183-2439. doi: 10.17645/mac.v9i3.3930. URL https://www.cogitatiopress.com/mediaandcommunication/article/view/3930.

[94] Eugenia Siapera. Multiculturalism online: The internet and the dilemmas of multicultural politics. *European Journal of Cultural Studies*, 9(1):5–24, February 2006. ISSN 1367-5494. doi: 10.1177/1367549406060804. URL https://doi.org/10.1177/1367549406060804.

[95] Luciano Floridi. The Informational Nature of Personal Identity. *Minds and Machines*, 21(4):549–566, November 2011. ISSN 1572-8641. doi: 10.1007/s11023-011-9259-6. URL https://doi.org/10.1007/s11023-011-9259-6.

[96] Adam Waytz, Nicholas Epley, and John T. Cacioppo. Social Cognition Unbound: Insights Into Anthropomorphism and Dehumanization. *Current Directions in Psychological Science*, 19(1):58–62, February 2010. ISSN 0963-7214, 1467-8721. doi: 10.1177/0963721409359302. URL http://journals.sagepub.com/doi/10.1177/0963721409359302.

[97] Laurel D. Riek, Tal-Chen Rabinowitch, Bhismadev Chakrabarti, and Peter Robinson. How anthropomorphism affects empathy toward robots. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, HRI '09, pages 245–246, New York, NY, USA, March 2009. Association for Computing Machinery. ISBN 978-1-60558-404-1. doi: 10.1145/1514095.1514158. URL https://doi.org/10.1145/1514095.1514158.

[98] Tony J. Prescott and Julie M. Robillard. Are friends electric? The benefits and risks of human-robot relationships. *iScience*, 24(1):101993, January 2021. ISSN 2589-0042. doi: 10.1016/j.isci.2020.101993. URL https://www.sciencedirect.com/science/article/pii/S2589004220311901.

[99] Christopher Burkett. I Call Alexa to the Stand: The Privacy Implications of Anthropomorphizing Virtual Assistants Accompanying Smart-Home Technology Notes. *Vanderbilt Journal of Entertainment & Technology Law*, 20(4):1181–1218, 2017. URL https://heinonline.org/HOL/P?h=hein.journals/vanep20&i=1229.

[100] Eloïse Zehnder, Jérôme Dinet, and François Charpillet. Anthropomorphism, privacy and security concerns: preliminary work. In *ERGO'IA 2021*, Bidart, France, October 2021. URL https://hal.archives-ouvertes.fr/hal-03365472.

[101] Bianca Kronemann, Hatice Kizgin, Nripendra Rana, and Yogesh K. Dwivedi. How AI encourages consumers to share their secrets? The role of anthropomorphism, personalisation, and privacy concerns and avenues for future research. *Spanish Journal of Marketing - ESIC*, January 2023. ISSN 2444-9709. doi: 10.1108/SJME-10-2022-0213. URL https://doi.org/10.1108/SJME-10-2022-0213.

[102] Andrew R. Chow. Why People Are Confessing Their Love For AI Chatbots. *Time*, February 2023. URL https://time.com/6257790/ai-chatbots-love/.

[103] Eneko Agirre and Aitor Soroa. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 33–41, Athens, Greece, March 2009. Association for Computational Linguistics. URL https://aclanthology.org/E09-1005.

[104] Sandra Wachter. Normative challenges of identification in the Internet of Things: Privacy, profiling, discrimination, and the GDPR. *Computer Law & Security Review*, 34(3):436–449, June 2018. ISSN 0267-3649. doi: 10.1016/j.clsr.2018.02.002. URL https://www.sciencedirect.com/science/article/pii/S0267364917303904.

[105] Kaan Varnali. Online behavioral advertising: An integrative review. *Journal of Marketing Communications*, 27(1):93–114, January 2021. ISSN 1352-7266. doi: 10.1080/13527266.2019.1630664. URL https://doi.org/10.1080/13527266.2019.1630664.

[106] Daniel Susser and Vincent Grimaldi. Measuring Automated Influence: Between Empirical Evidence and Ethical Values. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pages 242–253, New York, NY, USA, July 2021. Association for Computing Machinery. ISBN 978-1-4503-8473-5. doi: 10.1145/3461702.3462532. URL https://doi.org/10.1145/3461702.3462532.

[107] Xitong Guo, Yongqiang Sun, J. Yuan, Z. Yan, and N. Wang. Privacy-personalization paradox in adoption of mobile health service: The mediating role of trust. *Proceedings - Pacific Asia Conference on Information Systems, PACIS 2012*, January 2012.

[108] Stephen Armstrong. Data, data everywhere: the challenges of personalised medicine. *BMJ (Clinical research ed.)*, 359:j4546, October 2017. ISSN 0959-8138, 1756-1833. doi: 10.1136/bmj.j4546. URL https://www.bmj.com/content/359/bmj.j4546. tex.copyright: Published by the BMJ Publishing Group Limited. For permission to use (where not already granted under a licence) please go to http://group.bmj.com/group/rights-licensing/permissions.

[109] European Parliament. GDPR: 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016. URL https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng.

[110] Anita H. M. Cremers and Mark A. Neerincx. Personalisation Meets Accessibility: Towards the Design of Individual User Interfaces for All. In Christian Stary and Constantine Stephanidis, editors, *User-Centered Interaction Paradigms for Universal Access in the Information Society*, Lecture Notes in Computer Science, pages 119–124, Berlin, Heidelberg, 2004. Springer. ISBN 978-3-540-30111-0. doi: 10.1007/978-3-540-30111-0_9.

[111] Jeremy Knox, Yuchen Wang, and Michael Gallagher. Introduction: AI, Inclusion, and 'Everyone Learning Everything'. In Jeremy Knox, Yuchen Wang, and Michael Gallagher, editors, *Artificial Intelligence and Inclusive Education: Speculative Futures and Emerging Practices*, Perspectives on Rethinking and Reforming Education, pages 1–13. Springer, Singapore, 2019. ISBN 9789811381614. doi: 10.1007/978-981-13-8161-4_1. URL https://doi.org/10.1007/978-981-13-8161-4_1.

[112] Lucylynn Lizarondo, Saravana Kumar, Lisa Hyde, and Dawn Skidmore. Allied health assistants and what they do: A systematic review of the literature. *Journal of Multidisciplinary Healthcare*, 3:143–153, December 2010. ISSN null. doi: 10.2147/JMDH.S12106. URL https://www.tandfonline.com/doi/abs/10.2147/JMDH.S12106.

[113] Iason Gabriel. Artificial Intelligence, Values and Alignment. *Minds and Machines*, 30 (3):411–437, September 2020. ISSN 0924-6495, 1572-8641. doi: 10.1007/s11023-020-09539-2.

[114] Davit Marikyan, Savvas Papagiannidis, Omer F. Rana, Rajiv Ranjan, and Graham Morgan. "Alexa, let's talk about my productivity": The impact of digital assistants on work productivity. *Journal of Business Research*, 142:572–584, March 2022. ISSN 0148-2963. doi: 10.1016/j.jbusres.2022.01.015. URL https://www.sciencedirect.com/science/article/pii/S014829632200025X.

[115] Marguerita Lane and Anne Saint-Martin. The impact of Artificial Intelligence on the labour market: What do we know so far? Technical report, OECD, Paris, January 2021. URL https://www.oecd-ilibrary.org/social-issues-migration-health/the-impact-of-artificial-intelligence-on-the-labour-market_7c895724-en.

[116] Nicholas Crafts. Artificial intelligence as a general-purpose technology: an historical perspective. *Oxford Review of Economic Policy*, 37(3):521–536, September 2021. ISSN 0266-903X. doi: 10.1093/oxrep/grab012. URL https://doi.org/10.1093/oxrep/grab012.

[117] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*, 2023.

[118] Mustafa Suleyman. *The coming wave*. Crown, New York, first edition edition, 2023. ISBN 978-0-593-59396-7.

[119] Douglas Zytko, Pamela J. Wisniewski, Shion Guha, Eric P. S. Baumer, and Min Kyung Lee. Participatory Design of AI Systems: Opportunities and Challenges Across Diverse Users, Relationships, and Application Domains. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, pages 1–4, New York, NY, USA, April 2022. Association for Computing Machinery. ISBN 978-1-4503-9156-6. doi: 10.1145/3491101.3516506. URL https://doi.org/10.1145/3491101.3516506.

[120] Andrey Kormilitzin, Nenad Tomasev, Kevin R. McKee, and Dan W. Joyce. A participatory initiative to include LGBT+ voices in AI for mental health. *Nature Medicine*, 29(1):10–11, January 2023. ISSN 1546-170X. doi: 10.1038/s41591-022-02137-y. URL https://www.nature.com/articles/s41591-022-02137-y. tex.copyright: 2023 The Author(s), under exclusive licence to Springer Nature America, Inc.

[121] Rowena Cullen. Addressing the digital divide. *Online Information Review*, 25(5): 311–320, January 2001. ISSN 1468-4527. doi: 10.1108/14684520110410517. URL https://doi.org/10.1108/14684520110410517.

[122] Nick Couldry. Researching digital (dis)connection in the age of personalised media. In Graham Murdock and Peter Golding, editors, *Digital dynamics: engagements*

and connections., pages 105–124. Hampton Press Inc., Cresskill, NJ, 2010. ISBN 9731572739314.

[123] Elad Segev. *Google and the Digital Divide: The Bias of Online Knowledge*. Elsevier, January 2010. ISBN 978-1-78063-178-3.

[124] Christoph Lutz. Digital inequalities in the age of artificial intelligence and big data. *Human Behavior and Emerging Technologies*, 1(2):141–148, 2019. ISSN 2578-1863. doi: 10.1002/hbe2.140. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hbe2.140.

[125] Sophie Lythreatis, Sanjay Kumar Singh, and Abdul-Nasser El-Kassar. The digital divide: A review and future research agenda. *Technological Forecasting and Social Change*, 175:121359, February 2022. ISSN 0040-1625. doi: 10.1016/j.techfore.2021.121359. URL https://www.sciencedirect.com/science/article/pii/S0040162521007903.

[126] Vardhmaan Jain, Mahmoud Al Rifai, Michelle T. Lee, Ankur Kalra, Laura A. Petersen, Elizabeth M. Vaughan, Nathan D. Wong, Christie M. Ballantyne, and Salim S. Virani. Racial and Geographic Disparities in Internet Use in the U.S. Among Patients With Hypertension or Diabetes: Implications for Telehealth in the Era of COVID-19. *Diabetes Care*, 44(1):e15–e17, November 2020. ISSN 0149-5992. doi: 10.2337/dc20-2016. URL https://doi.org/10.2337/dc20-2016.

[127] Ophelia T. Morey. Digital Disparities. *Journal of Consumer Health on the Internet*, 11(4):23–41, December 2007. ISSN 1539-8285. doi: 10.1300/J381v11n04_03. URL https://doi.org/10.1300/J381v11n04_03.

[128] Eli Pariser. *The Filter Bubble: What The Internet Is Hiding From You*. Penguin UK, May 2011. ISBN 978-0-14-196992-3.

[129] Fabiana Zollo, Alessandro Bessi, Michela Del Vicario, Antonio Scala, Guido Caldarelli, Louis Shekhtman, Shlomo Havlin, and Walter Quattrociocchi. Debunking in a world of tribes. *PLOS ONE*, 12(7):e0181821, July 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0181821. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0181821.

[130] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118, March 2021. doi: 10.1073/pnas.2023301118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2023301118.

[131] Johanna Dunaway. Polarisation and misinformation. In *The Routledge Companion to Media Disinformation and Populism*. Routledge, 2021. ISBN 978-1-00-300443-1.

[132] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. Exposure to ideologically diverse news and opinion on Facebook. *Science (New York, N.Y.)*, 348(6239):1130–1132, June 2015. doi: 10.1126/science.aaa1160. URL https://www.science.org/doi/full/10.1126/science.aaa1160.

[133] Nathaniel Persily and Joshua A. Tucker. *Social Media and Democracy: The State of the Field, Prospects for Reform*. Cambridge University Press, September 2020. ISBN 978-1-108-85877-9.

[134] Jayson Harsin. Regimes of Posttruth, Postpolitics, and Attention Economies. *Communication, Culture and Critique*, 8(2):327–333, June 2015. ISSN 1753-9129. doi: 10.1111/cccr.12097. URL https://doi.org/10.1111/cccr.12097.

[135] Hunt Allcott and Matthew Gentzkow. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2):211–236, May 2017. ISSN 0895-3309. doi: 10.1257/jep.31.2.211. URL `https://pubs.aeaweb.org/doi/10.1257/jep.31.2.211`.

[136] Alexander Halavais. *Search Engine Society*. John Wiley & Sons, November 2017. ISBN 978-1-5095-1686-5.

[137] Ruha Benjamin. *Race After Technology: Abolitionist Tools for the New Jim Code*. John Wiley & Sons, July 2019. ISBN 978-1-5095-2643-7.

[138] Catharina O'Donnell and Eran Shor. "This is a political movement, friend": Why "incels" support violence. *The British Journal of Sociology*, 73(2):336–351, 2022. ISSN 1468-4446. doi: 10.1111/1468-4446.12923. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-4446.12923`.

[139] Kaitlyn Regehr. In(cel)doctrination: How technologically facilitated misogyny moves violence off screens and on to streets. *New Media & Society*, 24(1):138–155, January 2022. ISSN 1461-4448. doi: 10.1177/1461444820959019. URL `https://doi.org/10.1177/1461444820959019`.

[140] Petter Törnberg and Anton Törnberg. Inside a White Power echo chamber: Why fringe digital spaces are polarizing politics. *New Media & Society*, page 14614448221122915, September 2022. ISSN 1461-4448. doi: 10.1177/14614448221122915. URL `https://doi.org/10.1177/14614448221122915`.

[141] Matthew Gault. AI Trained on 4Chan Becomes 'Hate Speech Machine', June 2022. URL `https://www.vice.com/en/article/7k8zwx/ai-trained-on-4chan-becomes-hate-speech-machine`.

[142] Abhisek Tiwari, Sriparna Saha, Shubhashis Sengupta, Anutosh Maitra, Roshni Ramnani, and Pushpak Bhattacharyya. Persona or context? Towards building context adaptive personalized persuasive virtual sales assistant. In *Proceedings of the 2nd conference of the asia-pacific chapter of the association for computational linguistics and the 12th international joint conference on natural language processing (volume 1: Long papers)*, pages 1035–1047, Online only, November 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.aacl-main.76`.

[143] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5635–5649, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1566. URL `https://aclanthology.org/P19-1566`.

[144] Fajri Koto, Jey Han Lau, and Timothy Baldwin. Can Pretrained Language Models Generate Persuasive, Faithful, and Informative Ad Text for Product Descriptions? In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 234–243, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.ecnlp-1.27. URL `https://aclanthology.org/2022.ecnlp-1.27`.

[145] Daniel Susser. Invisible Influence: Artificial Intelligence and the Ethics of Adaptive Choice Architectures. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pages 403–408, New York, NY, USA, January 2019. Association for Computing Machinery. ISBN 978-1-4503-6324-2. doi: 10.1145/3306618.3314286. URL `https://doi.org/10.1145/3306618.3314286`.

[146] Ryan Calo. Digital Market Manipulation, August 2013. URL `https://papers.ssrn.com/abstract=2309703`. Place: Rochester, NY Type: SSRN Scholarly Paper.

[147] Daniel Susser, Beate Roessler, and Helen Nissenbaum. Online Manipulation: Hidden Influences in a Digital World, December 2018. URL https://papers.ssrn.com/abstract=3306006. Place: Rochester, NY Type: SSRN Scholarly Paper.

[148] Anthony Nadler, Matthew Crain, and Joan Donovan. The Political Perils of Online Ad Tech. Technical report, 2018. URL https://datasociety.net/wp-content/uploads/2018/10/DS_Digital_Influence_Machine.pdf.

[149] Morgan R. Frank, David Autor, James E. Bessen, Erik Brynjolfsson, Manuel Cebrian, David J. Deming, Maryann Feldman, Matthew Groh, José Lobo, Esteban Moro, Dashun Wang, Hyejin Youn, and Iyad Rahwan. Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences*, 116(14):6531–6539, April 2019. doi: 10.1073/pnas.1900949116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1900949116.

[150] Grace Lordan and David Neumark. People versus machines: The impact of minimum wages on automatable jobs. *Labour Economics*, 52:40–53, June 2018. ISSN 0927-5371. doi: 10.1016/j.labeco.2018.03.006. URL https://www.sciencedirect.com/science/article/pii/S0927537118300228.

[151] Mitch Downey. Partial automation and the technology-enabled deskilling of routine jobs. *Labour Economics*, 69:101973, April 2021. ISSN 0927-5371. doi: 10.1016/j.labeco.2021.101973. URL https://www.sciencedirect.com/science/article/pii/S0927537121000087.

[152] Jutta Haider, Malte Rödl, and Sofie Joosse. Algorithmically Embodied Emissions: The Environmental Harm of Everyday Life Information in Digital Culture, 2022. URL https://papers.ssrn.com/abstract=4112942. Place: Rochester, NY Type: SSRN Scholarly Paper.

[153] John Herrman. What Does It Mean That Elon Musk's New AI Chatbot Is 'Anti-Woke'? *New York Magazine*, November 2023. URL https://nymag.com/intelligencer/2023/11/elon-musks-grok-ai-bot-is-anti-woke-what-does-that-mean.html.

[154] Immanuel Kant. *Immanuel Kant: Groundwork of the Metaphysics of Morals: A German–English edition.* Cambridge University Press, 1 edition, February 2011. ISBN 978-0-521-51457-6 978-0-511-97374-1 978-1-107-61590-8. doi: 10.1017/CBO9780511973741. URL https://www.cambridge.org/core/product/identifier/9780511973741/type/book.

[155] B. Sharon Byrd and Joachim Hruschka. *Kant's Doctrine of Right: A Commentary.* Cambridge University Press, Cambridge, 2010. ISBN 978-0-521-19664-2. doi: 10.1017/CBO9780511712050. URL https://www.cambridge.org/core/books/kants-doctrine-of-right/5FBE982293DF392EC5C4950D57BD4BBE.

[156] John Rawls. *A Theory of Justice: Original Edition.* Harvard University Press, 1971. ISBN 978-0-674-88010-8. doi: 10.2307/j.ctvjf9z6v. URL https://www.jstor.org/stable/j.ctvjf9z6v.

[157] John Stuart Mill. *On Liberty.* Cambridge University Press, 1 edition, December 2011. ISBN 978-1-108-04083-9 978-1-139-14978-5. doi: 10.1017/CBO9781139149785. URL https://www.cambridge.org/core/product/identifier/9781139149785/type/book.

[158] Michael J. Sandel. *Liberalism and the Limits of Justice.* Cambridge University Press, Cambridge, 2 edition, 1998. ISBN 978-0-521-56298-0. doi: 10.1017/CBO9780511810152. URL https://www.cambridge.org/core/books/liberalism-and-the-limits-of-justice/6800BAC97E92FF5D64FF99DE858A900C.

[159] Jürgen Habermas, Thomas MacCarthy, and Jürgen Habermas. *Lifeworld and system: a critique of functionalist reason*. Number Vol. 2 in The theory of communicative action / Jürgen Habermas. Transl. by Thomas MacCarthy. Beacon, Boston, 1. digital-print ed edition, 2005. ISBN 978-0-8070-1401-1 978-0-8070-1400-4.

[160] Roger Clarke. Regulatory alternatives for AI. *Computer Law & Security Review*, 35 (4):398–409, August 2019. ISSN 0267-3649. doi: 10.1016/j.clsr.2019.04.008. URL https://www.sciencedirect.com/science/article/pii/S0267364919301281.

[161] Christiane Wendehorst. Strict Liability for AI and other Emerging Technologies. *Journal of European Tort Law*, 11(2):150–180, August 2020. ISSN 1868-9620. doi: 10.1515/jetl-2020-0140. URL https://www.degruyter.com/document/doi/10.1515/jetl-2020-0140/html?lang=en. Publisher: De Gruyter.

[162] Tarleton Gillespie, Patricia Aufderheide, Elinor Carmi, Ysabel Gerrard, Robert Gorwa, Ariadna Matamoros-Fernández, Sarah T. Roberts, Aram Sinnreich, and Sarah Myers West. Expanding the debate about content moderation: scholarly research agendas for the coming policy debates. *Internet Policy Review*, 9(4), October 2020. ISSN 2197-6775. URL https://policyreview.info/articles/analysis/expanding-debate-about-content-moderation-scholarly-research-agendas-coming-policy.

[163] Debra Satz. *Why Some Things Should Not Be for Sale: The Moral Limits of Markets*. Oxford University Press, May 2010. ISBN 978-0-19-987071-4. doi: 10.1093/acprof:oso/9780195311594.001.0001. URL https://academic.oup.com/book/1465.

[164] Ian Hosein, Prodromos Tsiavos, and Edgar A. Whitley. Regulating Architecture and Architectures of Regulation: Contributions from Information Systems. *International Review of Law, Computers & Technology*, 17(1):85–97, March 2003. ISSN 1360-0869, 1364-6885. doi: 10.1080/1360086032000063147. URL http://www.tandfonline.com/doi/full/10.1080/1360086032000063147.

[165] Lawrence Lessig. *Code: And Other Laws of Cyberspace*. ReadHowYouWant.com, July 2009. ISBN 978-1-4429-9637-3.