Journal of Artificial Intelligence Research 77 (2023) 377-394 Submitted 08/2022; published 05/2023

Your Prompt is My Command:
On Assessing the Human-Centred Generality of Multimodal Models

Wout Schellaert WSCHELL@VRAIN.UPV.ES
Fernando Martínez-Plumed FMARTINEZ@DSIC.UPV.ES
VRAIN, Universitat Politècnica de València, Spain

Karina Vold KARINA.VOLD@UTORONTO.CA
Institute for the History and Philosophy of Science and Technology
University of Toronto, Canada

John Burden JJB205@CAM.AC.UK
Leverhulme Centre for the Future of Intelligence
University of Cambridge, UK

Pablo A. M. Casares PABLOAMO@UCM.ES
Universidad Complutense de Madrid, Spain

Bao Sheng Loe A.LOE@JBS.CAM.AC.UK
Psychometrics Centre, University of Cambridge, UK

Roi Reichart ROIRI@TECHNION.AC.IL
Technion – Israel Institute of Technology, Israel

Sean Ó hÉigeartaigh SO348@CAM.AC.UK
Centre for the Study of Existential Risk, University of Cambridge, UK

Anna Korhonen ALK23@CAM.AC.UK
Language Technology Laboratory (LTL), University of Cambridge, UK

José Hernández-Orallo JORALLO@UPV.ES
VRAIN, Universitat Politècnica de València, Spain

Abstract

Even with obvious deficiencies, large prompt-commanded multimodal models are proving to be flexible cognitive tools representing an unprecedented generality. But the directness, diversity, and degree of user interaction create a distinctive "human-centred generality" (HCG), rather than a fully autonomous one. HCG implies that —for a specific user— a system is only as general as it is effective for the user's relevant tasks and their prevalent ways of prompting. A human-centred evaluation of general-purpose AI systems therefore needs to reflect the personal nature of interaction, tasks and cognition. We argue that the best way to understand these systems is as highly-coupled cognitive extenders, and to analyse the bidirectional cognitive adaptations between them and humans. In this paper, we give a formulation of HCG, as well as a high-level overview of the elements and trade-offs involved in the prompting process. We end the paper by outlining some essential research questions and suggestions for improving evaluation practices, which we envision as characteristic for the evaluation of general artificial intelligence in the future.

1. Introduction

A new paradigm of AI has emerged at the intersection of generative models and large language models. The resulting AI systems are able to perform a wide variety of tasks by being 'prompted', in which flexible inputs are 'continued' by equally flexible outputs. We introduce the term *massive multimodal models* (M*s) to emphasise both the connection with massive language models and their multimodal capabilities—both inputs and outputs contain snippets of, e.g., text, images, or audio, and the output modes may differ from the input modes[1]. New variants of these models are being released at a rapid pace, and we show an illustrative selection with different input/output modalities in Table 1.

Table 1: Illustrative selection of current *massive multimodal models* (M*s).

Modalities	**Models**
Text → Text	GPT3 (Brown et al., 2020), PaLM (Chowdhery & et al., 2022), BLOOM (BigScience et al., 2023), PanGu-α (Zeng et al., 2021)
Text → Image, or Text × Image → Image	DALL-E (Ramesh et al., 2021, 2022), Stable Diffusion (Rombach et al., 2022), Imagen (Saharia et al., 2022), Parti (Yu et al., 2022), MidJourney (Midjourney, 2022), GLIDE (Nichol et al., 2021)
Image × Text → Text	MAGMA (Eichenberg et al., 2022), Flamingo (Alayrac et al., 2022)
Speech → Speech	pGSLM (Kharitonov et al., 2022)
Text × Code → Code	Github Copilot/Codex (Chen et al., 2021)

Due to the flexibility of interaction and the reported versatility of the systems, interacting with prompt-commanded AI is different from other ways of interacting with machines, including other AI systems. This difference, together with the expectation of future availability and capability, demands a more systematic analysis of what 'prompting AI' implies for the evaluation of similar general systems. Concretely, we discuss what is new and how this affects human cognition (section 2), we consider the caveats of aggregation and the personal nature of cognition and utility (section 3), and we dissect the relevant elements of the prompting process (section 4). Lastly, we bolster our arguments by highlighting some active processes that effectually transform cognition (section 5), and finish with some essential research questions (section 6).

Takeaways

- M*s are best regarded as cognitive extenders, distinct from the autonomous perspective of AI systems. Their utility is therefore highly user-dependent.
- Properly evaluating them will require adopting methodologies from HCI.
- Feedback loops are emerging that impact evaluation and society at large.
- We provide actionable suggestions for tackling some of the resulting challenges.

1. We abstain from using the term "foundation" models (Bommasani et al., 2021), as we explicitly also include models not designed for fine-tuning.

2. A New Kind of Cognitive Tool

Cognitive tools are external artefacts that are used to aid the psychological capacities of the human brain in completing a cognitive task (Heersmink, 2021; Clark, 2008, 2004; Hutchins, 1999). Different tools place different cognitive needs upon users, either offloading or increasing particular cognitive demands (Gilbert et al., 2020; Risko & Gilbert, 2016; Sparrow et al., 2011; Clark, 2004). As with other technologies, the cognitive demands that M⋆s place on users will stem from their exact functional capacities and usage requirements. Purely illustrative, Fig. 1 shows an example of a cognitive task that is solved by a human using increasingly powerful cognitive artefacts.

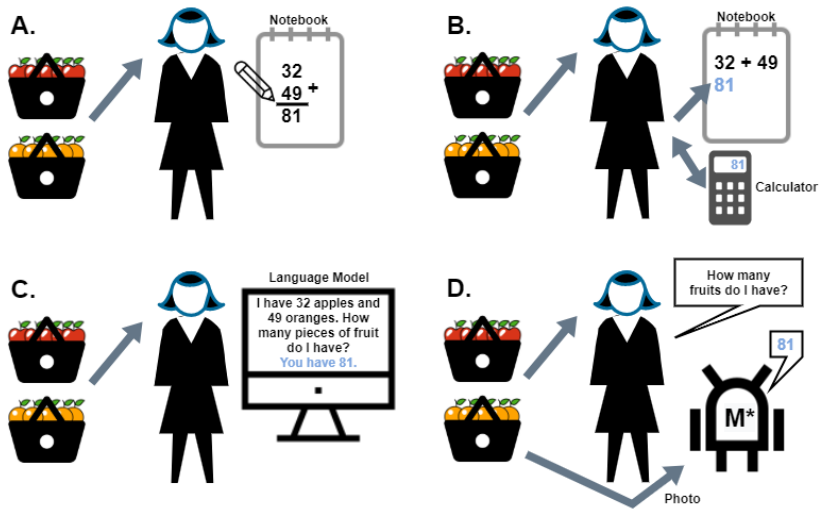

Figure 1: The evolution of cognitive extension for solving a grounded addition problem where Alice wants to know how many pieces of fruit she has in total. (A) Using a notebook. (B) Using a notebook and a calculator. (C) Using a language model. (D) Using a massive multimodal model M⋆. Ultimately Alice only needs to transform what she wants—knowing how many pieces of fruit there are in total—into an adequate multimodal prompt (a photo plus verbal question), thereby reducing the cognitive labour required of her.

There has been much discussion about the role of technology as cognitive extenders —tools that become a literal part of an agent's mind— in both philosophy and cognitive science, but this has largely focused on simple technologies, such as a pen and paper or a calculator, as shown in (A) and (B) in Fig. 1 (Clark & Chalmers, 1998; Menary, 2010). We argue that the combination of three properties makes the interaction with M⋆s unique: (i) *flexibility*, as in their input/output space, taking free-form language, images, code, etc.; (ii) *generality*, as they are applicable to a broad range of tasks; and (iii) *originality*, as they can be used to generate novel and original content. These features can be contrasted to the rigidity of other cognitive tools, such as to current digital assistants, which are restricted

in either task repertoire or searching the internet for content that must already exist, or to spelling correctors, which do generate original content.

Interactions between M⋆s and their users are also becoming more sophisticated than a single-step wrapping and unwrapping of multimodal elements. For example, Fig. 2 shows how GLIDE (Nichol et al., 2021) is used for step-by-step 'inpainting', an interactive generation where the output of the system is used as one of the inputs in the following iteration. Especially in these more complex interactions, the results obtained will depend on the effort and skills of the user when utilising the M⋆ to complete a cognitive task. For the user in Fig. 2 this involves figuring out the right order and prompts that condition the image generator in the direction they require.

Figure 2: Interactive generation with GLIDE. The first image is created from scratch using the first prompt, while subsequent 'inpainting' is generated using the previous image, the corresponding area marked in green, and the bottom prompt. Image taken from (Nichol et al., 2021).

The same features that make M⋆s unique also make attributions of success or failure more complicated. Both failure and success can be unexpected, and could relate to the phrasing of the prompt, the range of system capabilities, or the presence/absence of reference material in the training data. These attribution complexities are not something we would encounter with, for example, a calculator.

Hence, to properly evaluate these systems, the trade-offs users have to deal with need to be understood. These will include aspects such as the user's required cognitive effort, the probability of success, the intuitiveness of the system, or the reusability of a prompt. We discuss these in section 3 in a more complete overview of the involved elements. This analysis is common in the field of human-computer interaction (HCI) (Lazar et al., 2017; Rapp et al., 2021), especially for specific task-oriented tools, and even when using generative AI models cooperating with humans (Lee et al., 2022). However, they are only minimally considered when evaluating the generality of AI systems, as there is a pervasive notion that AI systems must be *autonomous* and 'generality' is often reduced to the notion of 'success in a wide range of tasks' (Legg & Hutter, 2007).

For evaluation of systems like M⋆s, used as cognitive *tools* or *extenders*, these agent-centric notions will have to give way.

3. Human-Centred Generality

Because of the directness, diversity, and degree of user interaction, AI systems such as M*s are 'human-centred' in a way that autonomous ones are not, and hence their generality must be understood and measured differently. While these systems fit perfectly in the new paradigm of 'human-centered AI' (Shneiderman, 2022), their generality and its evaluation have not been analysed in this context yet.

We say that a system displays human-centred generality (HCG) in so far as a user is able to use the system in (1) the completion of a wide range of cognitive tasks *relevant* to them, (2) with the commands that are *prevalent* to them and, (3) in a manner that is *effective* for them. For instance, if a system is capable of performing some tasks, but the user is not able to find the right command for any of them (in other words, if the desired behaviour is not easily accessible), we can say that the system is not general for this particular user.

HCG can be formulated as

$$V_h(\mathrm{M}^\star) = \sum_{t,p} \underbrace{\mathbb{P}(t|h)}_{Tasks} \cdot \underbrace{\mathbb{P}(p|t,h,\mathrm{M}^\star)}_{Prompts} \cdot \underbrace{v_h(\mathrm{M}^\star,t,p)}_{Utility}. \tag{1}$$

The expected utility of a particular M* for a particular user h is the sum of the utility for various tasks t and prompts p, weighted by how likely it is that the user requires the task (its relevance) and how likely it is that the user comes up with that particular 'prompt' for that task (its prevalence). The utility v_h is synonymous with the overall effectiveness of the M* for the task-prompt combination (also see section 4). Both distributions and the function v_h are complex components that need discussion, but it should be clear they are highly individual and may vary significantly for different pairs of $\langle h, \mathrm{M}^\star \rangle$. For example, it could be that some systems solve a wider range of tasks of interest for a user, covering $\mathbb{P}(t|h)$ better, but they may also require more pre- and post-processing from the user.

We also already know that some M*s work better —or are less harmful— for some individuals, groups or cultures (Abid et al., 2021; Bender et al., 2021; Hutchinson et al., 2020; Brown et al., 2020; Vig et al., 2020) and that direct negative impact can be caused by a lack of fairness or representation (Mehrabi et al., 2021; Cheng et al., 2021). In design and training of these systems, a social choice (Rossi et al., 2011) is implicitly made when aggregating preferences. The chosen interpretation of the user's prompt and the given output will be decided by existing inductive biases, which are a product of the perspectives represented in training data and system development. Our measure of human-centred generality would reflect the choices made and the biases that are present.

There are of course limits to how unique different people are; i.e., it can still make sense to use a fixed set of prompts and tasks, or to aggregate $V_h(\mathrm{M}^\star)$ for a particular distribution $p(h)$ over humans. But we hope that by formulating the definition like this, (i) it becomes explicit what distribution is chosen, e.g., for reasons of transparency, (ii) we pay attention to the differences across people and systems, and (iii) we can start zooming in on the cognitive costs involved, which is what we do in the next section.

4. Elements of the Prompting Process

In the generic prompting process, illustrated in Fig. 3, a user gets a task done by providing a multimodal 'prompt' to an M⋆, which responds, or 'continues' by outputting an answer. The result is then typically checked for adequacy and further transformed by the user before using it.

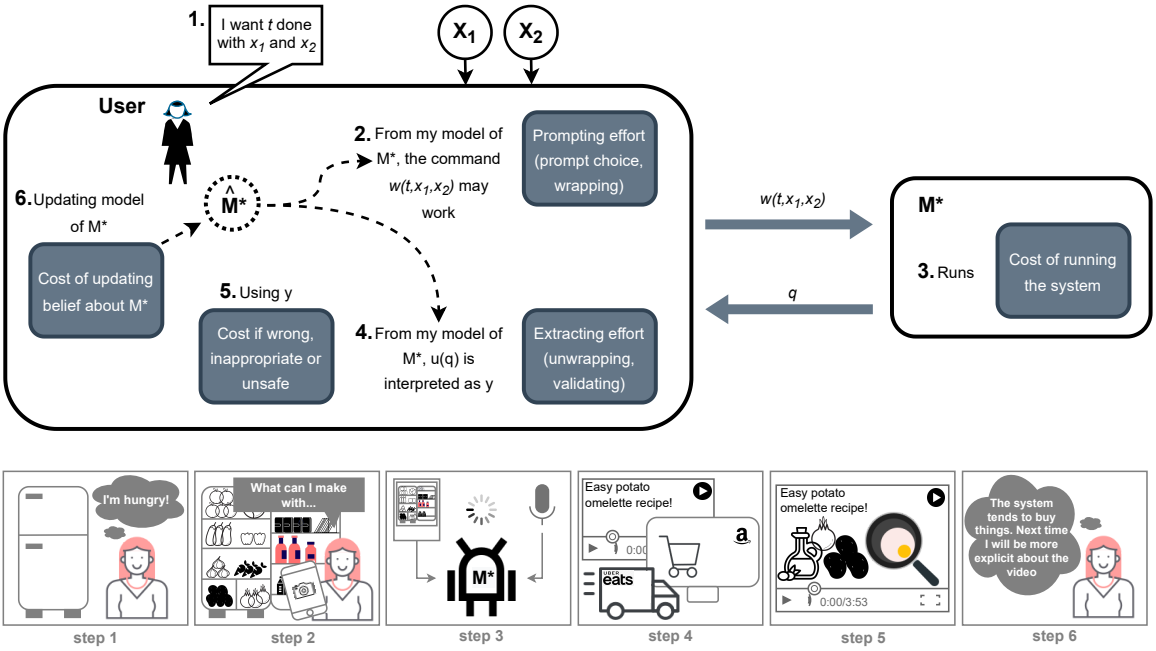

Figure 3: Top: Elements and process of a user (human) directly prompting an M⋆. The user wants a task t done (1) possibly including some multimodal elements x_1 and x_2 and, according to the internal model of M⋆ (denoted by $\hat{\text{M}}^\star$) has to articulate the wrapping of these elements $p = w(t, x_1, x_2)$ into a prompt for the system (2), producing an output q (3) from which the user unwraps the result $y = u(q)$, validates/assesses it (4) before finally using it (5). Given this iteration, the user updates (6) the internal model of M⋆. Bottom: Figurative six steps in the process of an M⋆ generating a recipe video from the contents of a user's refrigerator photo and a voice prompt. The user needs to deal with several outputs given by the M⋆, including not only a recipe video generated by M⋆ but also some other continuations such as buying more ingredients or ordering the whole dish to a restaurant, which she had to stop. After this interaction, the user updates her model of the system ($\hat{\text{M}}^\star$) thinking she should be more explicit about the recipe video next time.

Most parts of the process have received scientific attention independently: prompt choice and wrapping, e.g. (Liu et al., 2021; Reynolds & McDonell, 2021; Ben-David et al., 2022); improving the usefulness and safety of outputs, e.g., (Ouyang et al., 2022); or calibration, e.g. (Kumar, 2022; Lin et al., 2022; Kadavath et al., 2022).

But all of this effort is mostly made with regards to the correctness of the model, only in the subset of tasks where this is applicable, and assuming the prompts, transformations and other cognitive efforts are discounted or simply ignored. Nonetheless, the elements shown in Fig. 3 and further described in Table 2 paint a more complex picture than the error metrics used in standardised evaluation benchmarks of language models and generative models (Hendrycks et al., 2020; Brown et al., 2020; Reynolds & McDonell, 2021; Scao & Rush, 2021; Srivastava et al., 2022).

Benefit$^{(+)}$ or Cost$^{(-)}$	**Description**	**Origin**
Service$^{(+)}$	Potential gain of an answer such as a proof to a theorem, a generated image, auto-completed piece of code, the translation of some text, etc., which is relative to the problem and the cost of the user doing it themselves.	Task
Templating the prompt$^{(-)}$	Cost of devising a prompt template or schema, and anticipating the format of the result, according to the user's mental model of $\mathrm{M}^\star$, denoted by $\hat{\mathrm{M}}^\star$. For instance, the pairs `"Input:"` and `"Output:"` may be general prompts templates for many tasks, but are not sufficient for many others (Miltenberger, 2015; Laurel, 2013; Bourguet, 2003; Stivers & Sidnell, 2005).	System
Wrapping and unwrapping$^{(-)}$	A prompt can be reused for many instances, but then the relevant elements should be inserted along the prompt and extracted from it, as it is usually surrounded by irrelevant material, and both 'prompt and answer engineering' need to be anticipated (Liu et al., 2021).	System
Validation$^{(-)}$	Even if the user does not know the answer, for certain tasks they can at least validate if it is meaningful or fit for purpose (e.g., expecting a number when asking 'how many?').	Task
Consequences of incorrect or unsafe results$^{(-)}$	Cost measured in terms of the wide consequences of its use, including side effects or harmful stereotypes (Challen et al., 2019; Kocielnik et al., 2019; Russell et al., 2015; Venkatesh & Goyal, 2010; Vig et al., 2020).	Task
Miscalibration$^{(-)}$	Cost given the quality of the reliability or confidence given by the system, if so provided (Dinga et al., 2019), as it influences the validation process between steps 4 and 5 in Fig. 3.	System
User's training$^{(-)}$	Cost of building and updating the mental model $\hat{\mathrm{M}}^\star$ (Nelson & Cheney, 1987; Davis, 1989; DeLone & McLean, 1992) and the associated mental processes to interface with it more and more efficiently.	System
Reusability$^{(+)}$	The more diverse $\mathbb{P}(t\|h)$ is, the higher the (unit) cost in v_h and the lower the reusability will be. As in any automation or assistance problem, the number of tasks and instances that are repetitive compensate for the user's training costs, while also considering the times each prompt template can be reused according to task distribution.	Task

Table 2: Benefits and costs involved when using an $\mathrm{M}^\star$ as a cognitive tool. All of these should be considered as part of the utility function v_h in Eq. 1. The last column indicates whether the cost is mostly tied to the system or if the task also has a significant impact.

In general we would like to contrast the costs of the whole procedure against the utility of an acceptable answer. As shown in (Casares et al., 2022), the costs of interaction might outweigh the benefits, even when the system's outputs are of similar quality to those of the user themselves. The elements we lay out define the landscape of tasks that are most

suitable for M*s in terms of the cognitive costs and benefits involved. These trade-offs are intuitively considered when, for instance, our attention is captured by a generative system such as DALL-E (Ramesh et al., 2021) and its 'avocado chairs'[2]. But how useful is an M* for calculating the price of an item on sale? And how dangerous is it for determining the dose of a prescribed drug? While the benefit/cost breakout can help answer these questions in stable situations, users and M*s are also subject to continuous transformations, and the point in time the question is asked can have different answers due to these changes.

5. Bidirectional Cognitive Transformations

An additional complexity for robust evaluation, and an important topic for society at large, is that while these systems keep being adapted to humans, humans will simultaneously adapt themselves to these system as well. **We argue that M*s and their users will be continuously co-affecting one another in a self reinforcing *cognitive loop*.** This two-way interaction creates a cognitively coupled entity $\langle$human, M$^\star\rangle$ that could be evaluated as a cognitive system in its own right (Clark, 2008; Palermos, 2014). This critical feedback system is shown on the right-hand side of Fig. 4.

The original M*s are being adapted to the user, e.g. by giving more useful and factual answers through retrieval (Nakano et al., 2021; Borgeaud et al., 2022) or through fine-tuning with newly gathered human feedback and preferences (Ouyang et al., 2022; Askell et al., 2021). Meanwhile, humans are finding surprising new ways of making these systems work for their needs. Curious examples are adding the phrase "Yo be real" to a prompt to make the system indicate when it can not answer a question[3], or telling the system to "show their work" to improve arithmetic reasoning (Nye et al., 2021; Wei et al., 2022); entire guides[4] are being written for instructing these systems effectively, and even a 'prompt marketplace'[5] has popped up. In (Mishra et al., 2021) they find that adapting your instruction style has an impact that generalises across tasks. A more drastic adaption is the common need to switch to a non-native but well represented language (Wang et al., 2022), or the potential internalisation of systemic biases that are present in the output of these systems, e.g. regarding gender and professional occupation (Vig et al., 2020).

Throughout the prompting process, the user acts according to a mental model $\hat{\text{M}}^\star$ of the behaviour and capabilities of the actual system M*, and further interaction creates some meta-cognitive awareness of how best to interface with it. Transparency about system limitations could bootstrap this mental model. For example, most tools are upfront about the possible generation of harmful or biased content. Additionally, explainability and interpretability tools could help make the system more predictable quickly, but these are currently not implemented for the available public systems, and little XAI research has been done for systems at this scale. In any case the total utility $V_h(\text{M}^\star)$ and all of its elements are influenced greatly; as an example, this loop changes the range of tasks the user considers solvable.

2. These images are hard to create or find, and easy to unwrap and validate. https://openai.com/blog/dall-e/
3. https://www.gwern.net/GPT-3#expressing-uncertainty.
4. https://dallery.gallery/the-dalle-2-prompt-book/
5. https://promptbase.com/

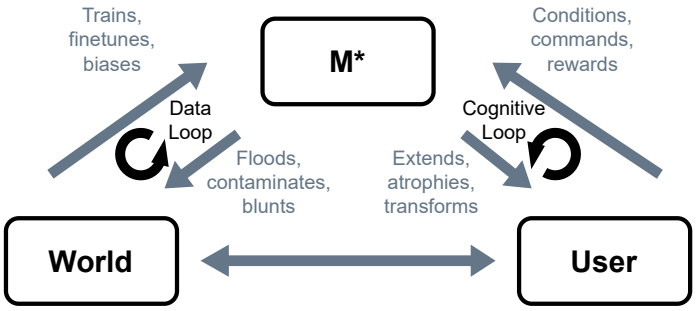

Figure 4: Interaction loops between humans (globally on the left and individually on the right) and M⋆s. Understanding the evolution of these systems will need to take these loops into account.

Moving to the left side of Fig. 5, we also identify a way in which M⋆s are dependent on interaction loops with external entities. Imagine that millions of images generated through a process like that depicted in Fig. 2 come to flood the Internet in the next few years, as is already happening with deepfakes (Fallis, 2021; Rini, 2020). Because M⋆s are commonly trained on web-collected data, which the M⋆ replicates (Bender et al., 2021), this interaction would create another feedback loop. **In the *data* loop, M⋆s are trained on existing web data and then create new data representing the bias and human distribution present in the source, which can then in turn be picked up for the training of new generations of systems**.

Transformation of human cognition has also happened with other related technologies, e.g., (Marsh & Rajaram, 2019; Ward et al., 2017; Applin & Fischer, 2015; Ferguson et al., 2015; Ward, 2013; Sparrow et al., 2011; Dror & Harnad, 2008; Woods & Hollnagel, 2006). We hypothesise that if these two loops are uncontrolled (i.e., left to evolve organically), they would cause negative consequences at both the individual and societal levels, e.g perpetuating existing social inequities (Bender et al., 2021). Instead, to control them, we need to better understand them and design appropriate interventions, e.g., by redesigning how these systems evolve and understanding how human cognition gets transformed with them. We collect some pertinent related research questions and suggestions in the following, concluding, section.

6. Forced Opportunities

The vision of commanding machines by *training* instead of *programming* them (Turing, 1950; Lieberman, 2001) is now leading way to machines that are *prompted*; unlike computer languages and data, multimedia prompts condition a *continuation*. This multimodal prompting is theoretically and practically unprecedented in AI, with only some related ideas having been hypothesised in early visions of human-computer interaction (Lieberman & Maulsby, 1996) or analysed at a philosophical level in a continuum of cognitive extension (Hernández-Orallo & Vold, 2019).

Holistic analysis of the benefits and efforts involved in the use of M⋆s Novel experimental analysis involving human interaction on the usefulness of language models can be found in (Casares et al., 2022) and (Lee et al., 2022), *but we suggest to extend this analysis to other modalities as well, and in general, to adopt these HCI-based methodologies more broadly.* Similar investigations like (Rapp et al., 2021) or (Schmidhuber et al., 2021) with chatbots could also serve as inspiration.
Discrepancies in generality across people and groups With the notion of human-centred generality in mind, we should give special care to any measured differences across people and groups. The costs of unfair discrimination, toxic language and harmful stereotypes quickly come to mind, but breakouts like Table 2 can highlight other nuanced discrepancies as well, *so we suggest to measure and compare them.* As discussed in (Weidinger et al., 2021)[Sec. 2], a system simply less well for you can be harmful too, e.g. by perpetuating social inequalities.
Prompt sensitivity and the use of 'promptese' languages M⋆s seem to be sensitive to the order of examples (Lu et al., 2022) and details in the wording that should be irrelevant, e.g. as in (Patel et al., 2021). *We suggest using multiple prompts and variations for what humans would consider 'the same question'.* This could help uncover whether the systems lack capability to complete the task, or the problem lies with specific language & its interpretation, which can easily vary between humans and cultures. For example, research like (Mishra et al., 2021) shows that adapting your instructional style can have an impact that generalises across tasks.
Cognitive loops, data loops, and their impact These prompt sensitivities might carry over to interaction with humans, and the same holds for any potential internalisation of what a normal system output is. We want to re-iterate that we should prevent our systems perpetuating stereotypes and exclusionary norms because they are present in the training data. We need to analyse how these systems transform users, cultures, and the data sources we use for training. For evaluation in particularly, we additionally care about train-test contamination (Brown et al., 2020)[Sec. 4], and distribution shifts (in users or data) (Ngo et al., 2021).
Formation of humans' mental models for various systems Especially related to cognitive loops, we have little insight into how humans form mental models of these M⋆s, how that influences HCG, what tasks they consider solvable, or how it compares to mental models for other AI systems. Given the non-use of explainability techniques in current systems, exploring ways for integration seems worthwhile to investigate as well. A concrete research question might be what the influence is of (un)wrapping assistance, i.e. techniques such as auto-prompting (Shin et al., 2020), prompt diversification (Jiang et al., 2020), or question decomposition (Perez et al., 2020)
Transparency on system limitations Transparency is an essential aspect of human-centred AI (European Commission, 2019), and we believe it to be valuable for both increasing human-centred generality (by bootstrapping users' mental-models), and for making human-centred evaluation more efficient (by reducing duplicate work and focusing evaluation efforts). Specifically, *we suggest including negative examples in published papers (or in linked repositories) and releasing logs of all prompts testers run against the system.* Additionally, a systems uncertainty about its output is a form of transparency (Bhatt et al., 2021). Some works investigate including e.g. textual notions of uncertainty in system output (Mielke et al., 2022; Lin et al., 2022), but this effort could be expanded to other modalities or other ways of integrating it into the user interfaces.

Table 3: Major research opportunities, challenges and suggestions for the evaluation of the generality of M⋆s and the consequence of their use. Especially concrete and actionable suggestions are *italicised*.

On on side, these systems are an ideal test bed for evaluating and steering general AI systems: they are the first incarnations of AI systems with this level of general utility and widespread usage, and many of the perspectives and challenges will carry over to more capable (and more impactful) systems. On the other hand, the issues we outline in the paper are issues worth tackling today. Indeed, "human-centred generality" forces us to rethink our evaluation procedures and borrow perspectives from the behavioural sciences while hybridising with those of artificial intelligence. In Table 3, we present a pertinent selection of these 'forced opportunities' as research topics interleaved with policy and methodology suggestions that might help tackle the evaluation and ethical challenges we raise.

In the norm, test batteries in AI lack human interaction (Hernández-Orallo et al., 2017; Shoham, 2017; Zhang et al., 2021; Martinez-Plumed et al., 2021). While still informative, and comparatively cheap, this user-agnostic benchmarking philosophy is insufficient. As in the area of HCI, we also need more holistic and realistic evaluations with humans to answer core questions. For which humans, tasks, and ways of interacting do these systems actually work? Why these humans, why those tasks? Since a cognitive tool can only be as general as it is effective for the users' relevant tasks and prevalent ways of interacting, our standard evaluation practices should aim to reflect this generality, any notable differences, and the social choices instilled into the tool that created them.

Acknowledgments

This work has been partially supported by the EU (FEDER) and Spanish MINECO grants RTI2018-094403-B-C32 and PID2021-122830OB-C42 (SFERA) funded by "ERDF A way of making Europe" and MCIN/AEI/10.13039/ 501100011033, Generalitat Valenciana under grant PROMETEO/2019/098, EU's Horizon 2020 research and innovation programme under grant agreement No. 952215 (TAILOR), the Future of Life Institute (FLI), under grant RFP2-152, and US DARPA HR00112120007 (RECoG-AI). This work was also funded by Estancias de Personal Investigador Doctor en Centros de Investigación Radicados fuera de la Comunitat Valenciana, "CIBEST/2021/30", the AI-Watch project by DG CONNECT and DG JRC of the European Commission, and the European Union with the "Programa Operativo del Fondo Europeo de Desarrollo Regional (FEDER) de la Comunitat Valenciana 2014-2020" under agreement INNEST/2021/317 (Neurocalçat) and by the Vic. Inv. of the Universitat Politècnica de València under "programa de ayudas a la formación de doctores en colaboración con empresas" (DOCEMPR21, DOCEMPR22). This work was also supported by the Social Sciences and Humanities Research Council of Canada's Insight Development Grant.

References

Abid, A., Farooqi, M., & Zou, J. (2021). Large language models associate Muslims with violence. *Nature Machine Intelligence*, *3*(6), 461–463.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A.,

Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., & Simonyan, K. (2022). Flamingo: A Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*.

Applin, S. A., & Fischer, M. D. (2015). New technologies and mixed-use convergence: How humans and algorithms are adapting to each other. In *2015 IEEE international symposium on technology and society (ISTAS)*, pp. 1–6. IEEE.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., & Kaplan, J. (2021). A General Language Assistant as a Laboratory for Alignment. *arXiv:2112.00861 [cs]*.

Ben-David, E., Oved, N., & Reichart, R. (2022). Pada: Example-based prompt learning for on-the-fly adaptation to unseen domains. *Transactions of the Association for Computational Linguistics*, *10*, 414–433.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). 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, p. 610–623, New York, NY, USA. Association for Computing Machinery.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., Nachman, L., Chunara, R., Srikumar, M., Weller, A., & Xiang, A. (2021). Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pp. 401–413, New York, NY, USA. Association for Computing Machinery.

BigScience, et al. (2023). BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv:2211.05100*.

Bommasani, R., et al. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G. B., Lespiau, J.-B., Damoc, B., Clark, A., et al. (2022). Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, pp. 2206–2240. PMLR.

Bourguet, M.-L. (2003). Designing and prototyping multimodal commands. In *Interact*, Vol. 3, pp. 717–720.

Brown, T., et al. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, Vol. 33, pp. 1877–1901. Curran Associates, Inc.

Casares, P. A. M., Loe, B. S., Burden, J., O'hEigeartaigh, S., & Hernandez-Orallo, J. (2022). How general-purpose is a language model? Usefulness and safety with human prompters in the wild. In *AAAI*.

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, *28*(3), 231–237.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Cheng, L., Varshney, K. R., & Liu, H. (2021). Socially Responsible AI Algorithms: Issues, Purposes, and Challenges. *Journal of Artificial Intelligence Research*, *71*, 1137–1181.

Chowdhery, A., & et al. (2022). PaLM: Scaling Language Modeling with Pathways. *arXiv:2204.02311 [cs]*.

Clark, A. (2004). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*. Oxford University Press.

Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Philosophy of Mind Series. Oxford University Press, New York.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7–19.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, pp. 319–340.

DeLone, W. H., & McLean, E. R. (1992). Information systems success: The quest for the dependent variable. *Information systems research*, *3*(1), 60–95.

Dinga, R., Penninx, B. W., Veltman, D. J., Schmaal, L., & Marquand, A. F. (2019). Beyond accuracy: measures for assessing machine learning models, pitfalls and guidelines. *bioRxiv*, p. 743138.

Dror, I. E., & Harnad, S. (2008). *Cognition distributed: How cognitive technology extends our minds*, Vol. 16. John Benjamins Publishing.

Eichenberg, C., Black, S., Weinbach, S., Parcalabescu, L., & Frank, A. (2022). MAGMA – Multimodal Augmentation of Generative Models through Adapter-based Finetuning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2416–2428, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

European Commission (2019). *Ethics Guidelines for Trustworthy AI*. Publications Office of the European Union, LU.

Fallis, D. (2021). The Epistemic Threat of Deepfakes. *Philosophy & Technology*, *34*(4), 623–643.

Ferguson, A. M., McLean, D., & Risko, E. F. (2015). Answers at your fingertips: Access to the internet influences willingness to answer questions. *Consciousness and Cognition*, *37*, 91–102.

Gilbert, S. J., Bird, A., Carpenter, J. M., Fleming, S. M., Sachdeva, C., & Tsai, P.-C. (2020). Optimal use of reminders: Metacognition, effort, and cognitive offloading.. *Journal of Experimental Psychology: General*, *149*(3), 501.

Heersmink, R. (2021). Varieties of artifacts: Embodied, perceptual, cognitive, and affective. *Topics in Cognitive Science*, *13*(4), 573–596.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Hernández-Orallo, J., Baroni, M., Bieger, J., Chmait, N., Dowe, D. L., Hofmann, K., Martínez-Plumed, F., Strannegård, C., & Thórisson, K. R. (2017). A new AI evaluation cosmos: Ready to play the game?. *AI Magazine*, *38*(3).

Hernández-Orallo, J., & Vold, K. (2019). AI extenders: the ethical and societal implications of humans cognitively extended by ai. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 507–513.

Hutchins, E. (1999). Cognitive artifacts. *The MIT encyclopedia of the cognitive sciences*, *126*, 127.

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5491–5501, Online. Association for Computational Linguistics.

Jiang, Z., Xu, F. F., Araki, J., & Neubig, G. (2020). How can we know what language models know?. *Transactions of the Association for Computational Linguistics*, *8*, 423–438.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Dodds, Z. H., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., & Kaplan, J. (2022). Language Models (Mostly) Know What They Know. arXiv preprint arXiv:2207.05221.

Kharitonov, E., Lee, A., Polyak, A., Adi, Y., Copet, J., Lakhotia, K., Nguyen, T. A., Riviere, M., Mohamed, A., Dupoux, E., & Hsu, W.-N. (2022). Text-free prosody-aware generative spoken language modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8666–8681, Dublin, Ireland. Association for Computational Linguistics.

Kocielnik, R., Amershi, S., & Bennett, P. N. (2019). Will you accept an imperfect AI? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–14.

Kumar, S. (2022). Answer-level Calibration for Free-form Multiple Choice Question Answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 665–679, Dublin, Ireland. Association for Computational Linguistics.

Laurel, B. (2013). *Computers as theatre*. Addison-Wesley.

Lazar, J., Feng, J. H., & Hochheiser, H. (2017). *Research methods in human-computer interaction*. Morgan Kaufmann.

Lee, M., Liang, P., & Yang, Q. (2022). Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *CHI Conference on Human Factors in Computing Systems*, pp. 1–19.

Legg, S., & Hutter, M. (2007). A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and applications*, *157*, 17.

Lieberman, H. (2001). *Your wish is my command: Programming by example*. Morgan Kaufmann.

Lieberman, H., & Maulsby, D. (1996). Instructible agents: Software that just keeps getting better. *IBM systems journal*, *35*(3.4), 539–556.

Lin, S., Hilton, J., & Evans, O. (2022). Teaching Models to Express Their Uncertainty in Words. arXiv preprint arXiv:2205.14334.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., & Stenetorp, P. (2022). Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Marsh, E. J., & Rajaram, S. (2019). The digital expansion of the mind: Implications of internet usage for memory and cognition. *Journal of Applied Research in Memory and Cognition*, *8*(1), 1–14.

Martinez-Plumed, F., Barredo, P., Heigeartaigh, S. O., & Hernandez-Orallo, J. (2021). Research community dynamics behind popular ai benchmarks. *Nature Machine Intelligence*, *3*(7), 581–589.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35.

Menary, R. (2010). *The extended mind*. Mit Press.

Midjourney (2022). Midjourney: text-to-art ai generator. `https://www.midjourney.com/`.

Mielke, S. J., Szlam, A., Dinan, E., & Boureau, Y.-L. (2022). Reducing conversational agents' overconfidence through linguistic calibration. arXiv preprint arXiv:2012.14983.

Miltenberger, R. G. (2015). *Behavior modification: Principles and procedures*. Cengage Learning.

Mishra, S., Khashabi, D., Baral, C., Choi, Y., & Hajishirzi, H. (2021). Reframing instructional prompts to gptk's language. *arXiv preprint arXiv:2109.07830*.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., et al. (2021). Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Nelson, R. R., & Cheney, P. H. (1987). Training end users: An exploratory study. *MIS quarterly*, pp. 547–559.

Ngo, H., Araújo, J. G., Hui, J., & Frosst, N. (2021). No news is good news: A critique of the one billion word benchmark. *arXiv preprint arXiv:2110.12609*.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., & Chen, M. (2021). GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., Sutton, C., & Odena, A. (2021). Show Your Work: Scratchpads for Intermediate Computation with Language Models. arXiv preprint arXiv:2112.00114.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Gray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Palermos, S. O. (2014). Loops, constitution, and cognitive extension. *Cognitive systems research*, *27*, 25–41.

Patel, A., Bhattamishra, S., & Goyal, N. (2021). Are NLP Models really able to Solve Simple Math Word Problems?. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094, Online. Association for Computational Linguistics.

Perez, E., Lewis, P., Yih, W.-t., Cho, K., & Kiela, D. (2020). Unsupervised question decomposition for question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8864–8880, Online. Association for Computational Linguistics.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR.

Rapp, A., Curti, L., & Boldi, A. (2021). The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, p. 102630.

Reynolds, L., & McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Rini, R. (2020). Deepfakes and the Epistemic Backstop. *Philosophers' Imprint*, *20*(24), 1–16.

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in cognitive sciences*, *20*(9), 676–688.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). Stable Diffusion. https://github.com/CompVis/stable-diffusion.

Rossi, F., Venable, K., & Walsh, T. (2011). *A Short Introduction to Preferences: Between Artificial Intelligence and Social Choice*, Vol. 5. Morgan & Claypool Publishers.

Russell, S., Dewey, D., Tegmar, M., Aguirre, A., Brynjolfsson, E., Calo, R., Dietterich, T., George, D., Hibbard, B., Hassabis, D., et al. (2015). Research priorities for robust and beneficial artificial intelligence. *The Future of Life Institute*.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*.

Scao, T. L., & Rush, A. M. (2021). How many data points is a prompt worth?. *arXiv preprint arXiv:2103.08493*.

Schmidhuber, J., Schlögl, S., & Ploder, C. (2021). Cognitive load and productivity implications in human-chatbot interaction. In *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*, pp. 1–6. IEEE.

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., & Singh, S. (2020). Eliciting knowledge from language models using automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4222–4235.

Shneiderman, B. (2022). *Human-Centered AI*. Oxford University Press.

Shoham, Y. (2017). Towards the AI index. *AI Magazine*, *38*(4), 71–77.

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, *333*(6043), 776–778.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Stivers, T., & Sidnell, J. (2005). Introduction: multimodal interaction. *Semiotica*.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*, 433–460.

Venkatesh, V., & Goyal, S. (2010). Expectation disconfirmation and technology adoption: polynomial modeling and response surface analysis. *MIS quarterly*, pp. 281–303.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., & Shieber, S. (2020). Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *NeurIPS*.

Wang, X., Ruder, S., & Neubig, G. (2022). Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 863–877, Dublin, Ireland. Association for Computational Linguistics.

Ward, A. F. (2013). Supernormal: How the internet is changing our memories and our minds. *Psychological Inquiry*, *24*(4), 341–348.

Ward, A. F., Duke, K., Gneezy, A., & Bos, M. W. (2017). Brain drain: The mere presence of one's own smartphone reduces available cognitive capacity. *Journal of the Association for Consumer Research*, *2*(2), 140–154.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G., & Gabriel, I. (2021). Ethical and social risks of harm from Language Models. *arXiv:2112.04359 [cs]*.

Woods, D. D., & Hollnagel, E. (2006). *Joint cognitive systems: Patterns in cognitive systems engineering*. CRC Press.

Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., et al. (2022). Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*.

Zeng, W., Ren, X., Su, T., Wang, H., Liao, Y., Wang, Z., Jiang, X., Yang, Z., Wang, K., Zhang, X., et al. (2021). Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. *arXiv preprint arXiv:2104.12369*.

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Lyons, T., Manyika, J., Niebles, J. C., Sellitto, M., et al. (2021). The AI index 2021 annual report. *arXiv preprint arXiv:2103.06312*.