Adobe and Meta Decry Misuse of User Studies in Computer Vision Research

Originally published at https://www.unite.ai/adobe-and-meta-decry-misuse-of-user-studies-in-computer-vision-research/

Adobe and Meta, together with the University of Washington, have published an extensive criticism regarding what they claim to be the growing misuse and abuse of user studies in computer vision (CV) research.

User studies were once typically limited to locals or students around the campus of one or more of the participating academic institutions, but have since migrated almost wholesale to online crowdsourcing platforms such as Amazon Mechanical Turk (AMT).

Among a wide gamut of grievances, the new paper contends that research projects are being pressured to produce studies by paper reviewers; are often formulating the studies badly; are commissioning studies where the logic of the project doesn't support this approach; and are often 'gamed' by cynical crowdworkers who 'figure out' the desired answers instead of really thinking about the problem.

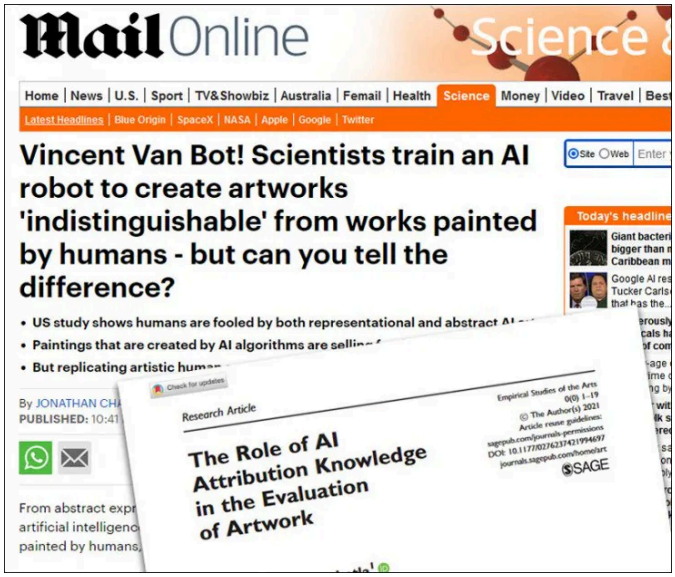
The fifteen-page treatise (titled *Towards Better User Studies in Computer Graphics and Vision*) that comprises the central body of the new paper levels many other criticisms at the way that crowdsourced user studies may actually be *impeding* the advance of computer vision sub-sectors, such as image recognition and image synthesis.

Though the paper addresses a much broader tranche of issues related to user studies, its strongest barbs are reserved for the way that *output evaluation* in user studies (i.e. when crowdsourced humans are paid in user studies to make value judgements on – for instance – the output of new image synthesis algorithms) may be negatively affecting the entire sector.

Let's take a look at a selection of some of the central points.

Sensational Interpretations

Among the paper's raft of suggestions for those who publish in the computer vision sector, is the admonition to 'interpret results carefully'. The paper cites one example from 2021, when a new research work claiming that 'individuals are unable to accurately identify AI-generated artwork' was widely spun in the popular press.



One of the higher-profile media reports on the 2021 paper 'The Role of AI Attribution Knowledge in the Evaluation of Artwork', by Harsha Gangadharbatla, cited as an example in the new paper. Here, The Daily Mail's source is The Times (paywalled). Sources: Daily Mail (archive link) / https://www.gwern.net/docs/ai/nn/gan/2021-gangadharbatla.pdf

The authors state*:

> *'[In] one study in a psychology journal, images of traditional artworks and images created by AI technologies were gathered from the web, and crowdworkers were asked to distinguish which images came from which sources. From the results it was concluded that "individuals are unable to accurately identify AI-generated artwork," a very broad conclusion that does not follow directly from the experiments.*
>
> *'Moreover, the paper does not report details about which specific image sets were collected or used, making the claims hard, if not impossible, to verify and reproduce.*
>
> *'More worrisome is that the popular press reported these results with the misleading claims that AIs can independently make art as well as humans.'*

Handling Crowdworkers Who Cheat

Crowdsourced workers are not usually paid much for their efforts. Since their prospects are minimal, and their best earning potential is through completing a high volume of tasks, many of them are, research suggests, disposed to take any 'shortcut' that will speed along the current task so that they can move on to the next minor 'gig'.

The paper observes that crowdsourced workers, much like machine learning systems, will learn repetitive patterns in the user studies that researchers formulate, and simply infer the 'correct' or 'desired' answer, rather than produce a true organic response to the material.

To this end, the paper recommends conducting checks on the crowdsourced workers, also known as 'validation trials' or 'sentinels' – effectively, fake sections of a test designed to see if the worker is paying attention, randomly clicking, or simply following a pattern that they have themselves inferred from the tests, rather than thinking about their choices.

The authors state:

> *'For instance, in the case of pairs of stylized images, one image of the pair can be an intentionally and objectively poor quality result. During analysis, data from participants that failed some preset number of the checks can be discarded, assumed to be generated by participants that were inattentive or inconsistent.*
>
> *'These checks should be randomly inserted in the study, and should appear the same as other trials; otherwise, participants may figure out which trials are the checks.'*

Handling Researchers Who Cheat

With or without intention, researchers can be complicit in this kind of 'gaming'; there are many ways for them, perhaps even inadvertently, to 'signal' their desired choices to crowdworkers.

For instance, the paper observes, by selecting crowdworkers with profiles that may be conducive to obtaining the 'ideal' answers in a study, nominally proving a hypothesis that might have failed on a less 'select' and more arbitrary group.

Phraseology is also a key concern:

> *'Wording should reflect the high-level goals, e.g., "which image contains fewer artifacts?" instead of "which image contains fewer color defects in the facial region?" Conversely, imprecise task wording leaves too much to interpretation, e.g., "which image is better?" may be understood as "which is more aesthetically-pleasing?" where the intention might have been to evaluate "which is more realistic?"*

Another way to 'benignly influence' participants is to let them know, overtly or implicitly, which of the possible choices in front of them is the author's method, rather than a prior method or random sample.

The paper states*:

> *'[The] participants may respond with the answers they think the researchers want, consciously or not, which is known as the "good subject effect". Do not label outputs with names like "our method" or "existing method". Participants can be biased by power dynamics (i.e., the researcher holding power by running the research session), researchers using language to prime participants (e.g., "how much do you like this tool that I built yesterday?"), and researchers and participants' relationship (e.g., if both work in the same lab or company).'*

The formatting of a task in a user study can likewise affect the neutrality of the study. The authors note that if, in a side-by-side presentation, the baseline is consistently positioned on the left (i.e. 'image A') and the output of the new algorithm on the right, study participants could infer that B is the 'best' choice, based on their growing presumption of the researchers' hoped-for outcome.

> *'Other presentation aspects such as the size of the images on the screen, their distance to each other, etc. may influence participant responses. Piloting the study with a few different settings may help spot these potential confounds early.'*

The Wrong People for the Wrong Product

The authors observe at several points in the paper that crowdsourced workers are a more 'generic' resource than would have been expected in previous decades, when researchers were forced to solicit help locally, often from faculty students who supplemented their income through study participation.

The requirement for active participation leaves the hired crowdworker little room to be 'nonplussed' by a product they are testing, and the paper's authors recommend that researchers *identify their target users* before developing and study-testing a potential product or service – else risk producing something very difficult to create, but that nobody actually wants.

> *'Indeed, we have often witnessed computer graphics or vision researchers attempting to get their research adopted by industry practitioners, only to find that the research does not address the target users' needs. Researchers who do not perform needfinding at the outset may be surprised to find that users have no need for or interest in the tool they have spent months or years developing.*
>
> *'Such tools may perform poorly in evaluation studies, as users may find that the technology produces unhelpful, irrelevant, or unexpected results.'*

The paper further observes that users who are actually likely to use a product should be selected for the studies, even if they are not easy to find (or, presumably, quite as cheap).

Rather than returning to recruiting on campus (which would be perhaps a rather backwards-looking move), the authors suggest that researchers 'recruit users in the wild', engaging with pertinent communities.

> *'For example, there may be a relevant active online message board or social media community that can be leveraged. Even meeting one member of the community may lead to* snowball sampling*, in which relevant users offer connections to similar individuals in their network.'*

Soliciting Feedback

The paper also recommends soliciting qualitative feedback from those who have participated in user studies, not least because this can potentially expose false assumptions on the part of the researchers.

> *'These may help debug the study, but they may also reveal unexpected facets of the output that influenced users' ratings. Was the participant "very unsatified" [sic] with the output because it was unrealistic, not aesthetic, biased, or for some other reason?*
>
> *'Without qualitative information, the researcher may work on refining the algorithm to be more realistic, instead of addressing the underlying user problem.'*

As with many of the recommendations throughout the paper, this particular recommendation involves further expenditure of time and money on the part of researchers, in a culture which, the work observes, is defaulting to rapid and practically obligatory crowdsourced user studies, which are usually fairly cheap, and which conform to an emerging study-driven culture that the paper criticizes throughout.

Over-Studied

The paper suggests that user studies are becoming a kind of 'minimum requirement' in the pre-print computer vision community, even in cases where a study cannot be reasonably formulated (for instance, with an idea so novel or marginal that there is no 'like-for-like' analysis to conduct, and which may not be susceptible to any reasonable metric that could yield meaningful results in a user study).

As an example of 'study bullying' (not the authors' phrase), the researchers cite the case of an ICLR 2022 paper for which peer reviews are available online (archive snapshot taken 24th June 2022; link taken directly from the new paper)†:

> *'Two reviewers gave very negative scores due, in part, to a lack of user studies. The paper was eventually accepted, accompanied by a summary chastizing the reviewers for using "user studies" as an excuse for poor reviewing, and accusing them of gatekeeping. The full discussion is worth reading.*
>
> *'The final decision noted that the submission described a software library that* ***had been deployed for years****, with thousands of users (information that was not revealed to the reviewers for anonymous review). Would the paper—which describes a highly impactful system—have been rejected if the committee had not had this information?*
>
> *'And, had the authors gone through the extra effort of contriving and performing a user study, would it have been meaningful, and would it have been enough to convince the reviewers?'*

The authors state they have seen reviewers and editors impose 'onerous evaluation requirements' on submitted papers, notwithstanding whether such evaluations would really have any meaning or value.

> *'We have also observed authors and reviewers use MTurk evaluations as a crutch to avoid making hard decisions. Reviewer comments like "I can't tell if the images are better, maybe a user study would help" are potentially harmful, encouraging authors to perform extra work that will not improve a lackluster paper.'*

The authors close the paper with a central 'call to action', for the computer vision and computer graphics communities to consider more fully their requests for user studies, instead of letting a study-driven culture develop as a rote default, notwithstanding the 'edge cases' where some of the most interesting work may not fit some of the most profitable or

fruitful research and submission pipelines.

The authors conclude:

> *'[If] the primary goal of running user studies is to appease reviewers rather than to generate new learnings, the utility and validity of such user studies should be put into question by authors and reviewers alike. Penalizing work that does not contain user evaluation has the unintended consequence of incentivizing hastily done, poorly executed user research.*
>
> *'A maxim to keep in mind is that "bad user research leads to bad outcomes", and such research will continue if reviewers continue to ask for it.'*

** My conversion of the paper's inline citations to pertinent hyperlinks*

† *My **emphasis**, not the authors'.*

First published 24th June 2022.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More