Does ChatGPT mean we have to change how we assess?

Daisy Christodoulou

ChatGPT is capable of producing original and high-quality essays with minimal effort.

What does this mean for educational assessment?
Many people argue that it shows the paucity of our current assessment tasks. If we are setting assessments that a robot can complete, surely that shows the assessments are not good enough or hard enough or just plain 'human' enough and that they therefore need reworking.

I have previously appeared on Good Morning Britain where Dan Fitzpatrick made exactly this argument, and Susannah Reid agreed with what she called a 'profound' point. (Unfortunately, the debate ended there, and I did not get the chance to respond.)

They are not the only ones. In the *Financial Times*, the columnist Camilla Cavendish suggested the following: "Rather than banning ChatGPT, teachers should ask pupils to give it an assignment and critique its response."

Marc Andreessen has weighed in with a similar argument on X. "'ChatGPT plagiarism' is a complete non-issue. If you can't out-write a machine, what are you doing writing?"

And even before ChatGPT existed, people were making this argument in the context of other technologies. Here is the economist Daniel Susskind in his book A World Without Work.

> Think of the way that we teach and test mathematics, for instance. Many of the problems we set students in secondary school, if not university, can now be solved by apps like PhotoMath and Socratic: take a photo of the problem, printed or handwritten, with a smartphone, and these apps will scan it, interpret it, and give you an instant answer. It is not a good sign that we still teach and test mathematical material in such a routine way that free off-the-shelf systems like these can handle lots of it with ease.

As well as those examples, I have heard similar sentiments expressed to me by many, many people over the last month or so, to the extent that it almost seems as though this is the prevailing opinion. ChatGPT can write essays? We will have to set harder/different essays or assessments then!

As popular as this argument is, I disagree with it. There are three big flaws with the 'we should set assessments that computers can't complete' argument. It's an argument that misunderstands basic principles about (1) technology, (2) education and (3) assessment.

1 It's quite hard to find educational assessments computers can't do
Even before ChatGPT, this was true, as Daniel Susskind's point inadvertently makes clear. Now ChatGPT has come along: as well as writing essays, it can get passing marks in a number of prestigious professional qualifications. The solution of 'getting kids to critique a ChatGPT essay' is not going to work either,

as ChatGPT is rather good at critiquing its own responses. I suspect it would also be good at critiquing critiques of its responses, and critiquing critiques of those, and so on *ad infinitum*.

The basic technological principle here is Moravec's paradox, first developed in 1988, which is that computers find the types of academic skills we teach and assess in schools trivially easy. In Moravec's (1988) words: "It is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility" (p. 15).[1]

Hans Moravec suggested a biological basis for his paradox: humans as a species have been evolving visual and spatial skills for millions of years, but abstract thought only for about a hundred thousand or so.

OK, you might say. This is all very interesting, but doesn't it just prove the point that we need to change the way we teach and assess? If computers really are so brilliant at these typical academic skills that are taught in schools, maybe we should stop teaching them completely or only teach the particularly advanced, specialist and niche ones that computers can't do?

No. First of all, we will always want to teach academic skills for personal development. It's good to be able to read, write and count, even if a computer is faster and quicker. We didn't stop teaching PE because of the invention of the car or drawing because of the invention of the camera.

It is true that in order to develop, understand and use advanced technologies, humans are going to need

advanced skills, and it is true that these more advanced skills should be one of the *ultimate* aims of education.

But this does not mean that we can forget about the more fundamental skills, because they are what allow us to develop the more advanced skills. This brings us to the second flaw in the 'set more complex assessments' argument.

2 If we want students to have advanced skills, they can't leapfrog fundamental skills

Of course, we want our students to develop the higher order skills of being able to critique writing produced by AI chatbots and to direct the outputs of new technologies. But those skills *depend* on more fundamental skills and there is no way we can jump ahead to the more advanced skills without acquiring the more basic skills first. In order for students to successfully grapple with problems computers *cannot* do, they must work through problems that computers *can* do. If schools could only teach maths using problems that computers cannot solve, we would have to teach six-year-olds maths using problems even top mathematicians find difficult!

So, it doesn't matter if we set our students tasks that can be easily solved by computers. It doesn't matter if they produce writing that is weaker than that of ChatGPT. The easy problems and the weak writing are milestones on their journey to mastery which cannot be skipped or outsourced.

The basic educational principle here is to do with the relationship between working memory and long-term memory. We have limited working memories, so we need to make up for that weakness by storing lots of information in long-

We will always want to teach academic skills for personal development. It's good to be able to read, write and count, even if a computer is faster and quicker.

term memory. You can't outsource that information to Google or ChatGPT: it needs to be in long-term memory so it can be effortlessly and frictionlessly summoned to working memory when needed, and combined with information in the environment, where it will produce what we typically call 'skill'. Here's a quotation from Daniel Willingham, Professor of Psychology at the University of Virginia, which expresses this well.

---

[1] Of course computers are getting better at perception and mobility – but Moravec's paradox still holds in that the computers are requiring a lot more computing power to achieve them.

Does ChatGPT mean we have to change how we assess?

*Data from the last 40 years lead to a conclusion that is not scientifically challengeable: thinking well requires knowing facts, and that's true not simply because you need something to think about. The very processes that teachers care about most – critical thinking processes such as reasoning and problem solving – are intimately intertwined with factual knowledge that is stored in long-term memory (not just found in the environment). (*Willingham, 2021*, p. 28)*

There is an analogy here with chess. Chess computers have been better than the very best humans at chess for decades now. What do we do if a child wants to learn chess? Do we say, well, there is no value in teaching or assessing any content that a chess machine can do? Do we say, we need to set them problems that Alpha Go cannot solve? Of course not! We teach them how the pieces move, what the basic openings are, and some of the common patterns to look

out for, even though computers find all these tasks trivially easy. Interestingly, we can and do use technology to help students *acquire* these basics, but at no point do we assume technology means they *never have to learn* those basics. The same is true of other skills.

3 The point of an assessment is not the product but the process

The value of the work students produce in an assessment is not in the work itself but in understanding it represents and the thinking that went into creating it. Imagine two students write an essay. One struggles hard, reads a lot, writes and redrafts it, and produces something that is OK but not great. Another produces something perfect by pasting the prompt into ChatGPT and copying the output into a Word doc. Who has done better? If all we cared about was the product, then it would be the second student. But we don't. We care about the process. The first student's response has led to them learning more and their essay represents greater understanding of the topic than the second student's essay. Fundamentally, it does not matter that a computer can answer this question better than the student. What matters is what the student has learned from answering it, and what it tells us about their understanding.

The basic assessment principle here is the difference between the sample and the domain. The sample is the test itself. The domain is the student's wider understanding. The sample only matters if it tells us something valuable about the domain – otherwise it is worthless. This might seem like a fairly straightforward distinction, but even before ChatGPT it was widely misunderstood. Here's what Daniel Koretz, Professor of Educational Assessment at Harvard, has to say about it.

*This might be called the sampling principle of testing: test scores reflect a small sample of behaviour and are valuable only insofar as they support conclusions about the larger domains of interest. This is perhaps the most fundamental principle of achievement testing. A failure to grasp this principle is at the root of widespread misunderstandings of test scores. (*Koretz, 2008*, pp. 21–22)*

The reaction to ChatGPT bears out Koretz's point about how poorly understood this principle is. I also think this misunderstanding could cause real problems with student motivation. If a student struggles for an hour over an extended piece of writing and then finds that a computer has surpassed it in seconds, it is entirely possible they will feel demotivated. What they need to hear from adults is "Don't worry, your work is of value, you're on a journey and you are developing your own writing skills." What they don't need to hear is "Well there's no point in even bothering, the computer is so much better than you. Try this assessment which is even harder instead!"

So, if we are setting assessments that a robot can complete, what does that say about our assessments? It doesn't tell us very much at all. Maybe it's a good assessment, maybe it's not. Whether a robot can complete it or not is largely irrelevant when judging its quality.

The one way in which the answer to this question will matter is in terms of the conditions that the assessment should be taken in, something I'll consider in future articles.

This article originally appeared on the No More Marking Substack. Subscribe for more updates: https://substack.nomoremarking.com/subscribe.

Daisy Christodoulou [@daisychristo on X] is Director of Education at No More Marking, a provider of online Comparative Judgement software for schools. She is also the author of three books: Teachers vs Tech, Making Good Progress?, and Seven Myths about Education. In 2019 Daisy was recognised in the New Year's Honours List with an MBE for services to education.