AI isn’t always bad news. Here’s how to deploy it ethically.

It’s fair to say that AI gets a bad rep. It’s a powerful tool: and when used incorrectly or irresponsibly, can have far-reaching negative consequences. In the last few months alone, tech behemoths Microsoft and Meta have retracted AI models following legitimate safety and inaccuracy fears. But what if instead of simply retracting models because they are unsafe, these companies cleaned up their act and created ethical AI models from the get-go that could be used more safely?

The definition of ethical AI is in constant evolution – and we certainly have not yet found a perfect version. Despite this, deploying AI in a responsible way shouldn’t be so hard. In fact, it should be standard best practice to mitigate risks, prevent harm and enable businesses to deploy AI with confidence.

Yet despite *some progress* in creating a framework for using and developing ethical AI, we don’t yet have any AI-specific regulation from the UK government. This can make it difficult for companies to know where to start.

I’m a theoretical physicist and have worked in AI and machine learning for over a decade. I am currently CTO at thymia, where I build ethical AI models for use in a health context. Here’s my advice.

Build ethics and transparency into your AI models from day one

It can be difficult to retrofit ethics into an AI model once it has already been built and calibrated on a particular dataset. Microsoft recently released its ‘Responsible AI Standard’ which is why it then had to retract some of its AI products: the products already created were not in line with these new objectives. To avoid this kind of mess, it is critical to set clear principles from the outset that will inform the ethical design and development of your AI.

AI models are only as accurate and ethical as the data they’re trained on, so it’s important to aim to train your AI systems on clean and debiased datasets as much as possible.

That having been said, the reality is that when you’re dealing with human data, some bias will often remain, despite your best efforts in collecting and curating it cleanly.

The ethical response to this problem is transparency. When you’re first gathering data to calibrate your models, ask yourself the following: are the people you’re gathering data from aware of what is being collected, how it will be stored and processed, and what it will be used for? And as you onboard new users and their data feeds into the system, are these people aware of how your tool is generating its output, how to interpret it, and what limitations it encodes? Finally, are the people interpreting your system’s outputs also aware of its limitations?

Always be clear and transparent about your model’s limitations and ensure humans are involved in the communication loop at every step, so that users understand an AI’s decision is not always final, and that other factors could be at play.

You should also commit to taking data privacy and security seriously, ensuring your systems can stand up to attack, are not susceptible to leaks or breaches and follow all relevant rules and regulations in your region. For example, if you’re claiming to provide ‘ethical’ AI services in mental health, it’s critical to have cybersecurity certifications, such as ISO27001, and – if in the U.S. – HIPAA compliance. In Europe, you should be GDPR-compliant by default. Not having these certifications would be a serious red flag.

Do not overstate the success of your tool

Another very important red flag to look out for when you're building and testing ethical AI models is producing a surprisingly high quality score for a model that's performing a difficult task – particularly if your training dataset is small and you are not able to offer clarity on your testing methodology.

In emerging fields (such as the one I am currently in: using AI to identify biomarkers from voice and video) there is a troubling tendency to *overstate results in small scale studies*. You see claims of above 90% predictive accuracy in studies that gathered data from perhaps 100 people. But releasing inadequately backed statements with poor methodology is misleading and risks damaging the reputation of the field of AI as a whole. This malpractice has led to worrying statements from important industry voices such as the Information Commissioner's Office, who *recently released a statement* warning companies against using 'immature' AI technologies due to the potential risks.

By testing and reporting on the success of new AI tools accurately, we can start to restore the credibility of AI technologies and showcase the incredible scientific progress happening within the field of ethical AI; which, when used in context and with awareness of its limitations, has immense value and potential. Blanket limitations on the use of AI by default do not just remove poor AI models from the market, they also simultaneously remove any potential for positive impact from well-trained, scientifically-backed and ethical AI models as well.

Scientifically validate your models

When you're building something like a medical device, it's subject to rigorous scientific testing and regulatory procedures to ensure it is safe to use. But when building a wellness product not classified as a 'medical device', or any other kind of AI-powered tech product, it often won't be subject to the same level of scrutiny. This is alarming, especially when many such AI tools are going directly into the hands of ordinary people, without rigorous regulation or testing.

One way to overcome the natural distrust this can cause is to scientifically validate the claims you are making about what your AI models can and cannot do, by publishing your results in a peer-reviewed journal or conference paper. This is the only way to ensure your tool is criticised and put under the scrutiny of people with the appropriate level of knowledge in order to ensure its quality, even if your product isn't covered by regulation. This type of peer review

signals to users that your models are robust and can be trusted.

When building AI for health applications, secure clinical validation too

At thymia, we're building AI tools for clinicians to use to better assess mental health. So it was critical that we involved clinicians in the development and evaluation of our tool. What's more, especially when first entering the market, we made sure that we were not the last port of call for deciding on a particular diagnosis or treatment, but that a clinician used their own expert knowledge alongside our model outputs and more traditional methods. If you are building AI for clinical settings, clinical validation and input is essential in order to ensure your tool is safe to use and that patients are being presented with outputs in a safe, professional and empathetic way, in the presence of a clinician who can help them make sense of it.

If you're building a direct-to-patient health tool, such as a B2C health app, it becomes even more important to proceed with care and to seek medical input when you assess how to present outputs to users. Even seemingly harmless outputs, such as those of a mood or fatigue tracker, can be dangerous if presented in the wrong way without a clinician being present to individuals with conditions such as depression.

AI is a complex and powerful beast. It is not inherently good or bad. But until better regulation is in place, rightly or wrongly, it is up to individual companies and developers to hold themselves to the highest standards and to create AI systems responsibly and ethically. For advice and guidance, *Holistic AI* is a brilliant source of information and expertise. Only by holding ourselves accountable and deploying AI in an ethical way will we be able to overcome the distrust surrounding the technology, allay safety fears and enable its most useful, valuable and life-enhancing applications to shine through.

Stefano Goria is CTO and co-founder at *thymia.*

Article by STEFANO GORIA