What is AI?

A Guide

Media Smarts

Meta

GUIDE

WHAT IS AI?

You've probably heard the phrase "AI" before. You might have seen news stories about the amazing things that it can do, like creating images or videos in minutes or carrying on what seem like real human conversations. AI has been described as a "game changer"[1] for people with disabilities, making it possible to automate tasks that otherwise would have been hugely time-consuming. AI chatbots have also been found to make people less lonely.[2] But what is AI, exactly, and what issues should I be watching out for? What are the benefits and what are the risks?

This guide provides an overview of what AI is – and in particular Generative AI – and gives two examples of main AI tools you are likely to encounter. Then it explains some key ethical and social issues related to Generative AI.

What exactly is AI?

AI (artificial intelligence) is a way of using *computer algorithms* to do things with little or no human involvement.

An **algorithm** is basically a series of steps or instructions for doing something. AI algorithms aren't programmed but *trained*. This means they're given a data set to learn from, such as a collection of millions of pictures or written texts. They find patterns or connections in the data set and use those to solve the problem they've been programmed to solve.

"You don't need to produce a precise list of instructions and communicate them... You give the machine data, a goal and feedback when it's on the right track – and leave it to work out the best way of achieving the end." Hannah Fry, Hello World

AI algorithms are much more powerful and flexible than algorithms that are written by humans, but they're also harder to analyze and to understand. While we might know what data goes into the algorithm, and we can see what is produced by them, we can't easily tell the *process* between the two. This is why AI is sometimes referred to as a "black box." Because the most sophisticated ones are able to change and adapt over time, even the people who make and operate them may not fully know how they work.

Even though AI isn't programmed in the traditional way, people are still a necessary piece of the training process. They give feedback by rating the quality of

1 Aquino, S.(2024) AI could be a game changer for people with disabilities. MIT Technology Review. https://www.technologyreview.com/2024/08/23/1096607/ai-people-with-disabilities-accessibility

2 De Freitas, J., Uguralp, A. K., Uguralp, Z. O., & Stefano, P. (2024). AI Companions Reduce Loneliness. *arXiv preprint arXiv:2407.19096*.

answers, captioning or annotating things in the data set, or testing to make sure that it doesn't produce graphic, violent or other inappropriate content.[3] AI "seems so human because it was trained by an AI that was mimicking humans who were rating an AI that was mimicking humans who were pretending to be a better version of an AI that was trained on human writing."[4]

What is Generative AI?

Generative AI is what we call AI systems that can *generate* things like images, video, voice and text. They do this by first *encoding* many examples of the kind of content they're going to make, and then *decoding* to make something new.[5]

From the user's point of view, using generative AI starts with providing what's called a *prompt*: a description of what you want the AI to generate (a text, image, video, piece of music, et cetera.) These can be very simple (such as "an apple") or may include instructions about *how* to make it (such as "an apple in the pointillistic style of Cezanne"). Prompts can also put limits on what the AI can do or ask it to take on a particular role.

Let's look at the two most common examples of generative AI:

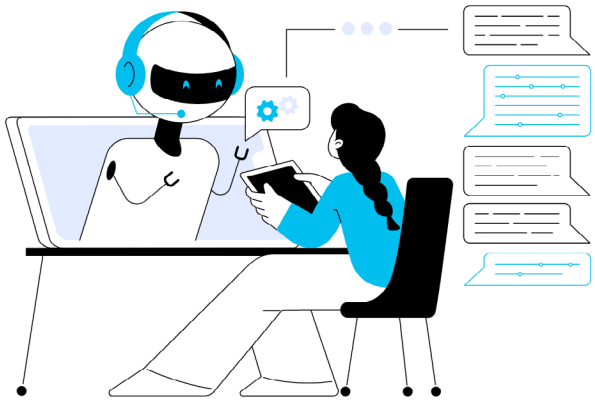
CHATBOTS

Chatbots, which can produce written text, answer questions, and even carry on conversations, are based on a kind of AI called a **large language model**.

What does that mean? Let's go through the three words in reverse order.

Model: Like other machine learning algorithms, most of what chatbots can do isn't programmed, but instead come from being trained on large amounts of writing. They find patterns in these to create a model of how language works.

Language: Chatbots can read and write fluently at the level of sentences, paragraphs and even full articles. They do this mostly by using what are called *transformers* to look at how similar or different words



3 Hao, K., & Seetharaman D. (2023) Cleaning Up ChatGPT Takes Heavy Toll on Human Workers. *The Wall Street Journal.*

4 Dzieza, J. (2023) AI Is A Lot of Work. *New York.* https://nymag.com/intelligencer/article/ai-artificial-intelligence-humans-technology-business-factory.html

5 Murgia, M. (2023) Transformers: the Google scientists who pioneered an AI revolution. Financial Times. https://www.ft.com/content/37bb01af-ee46-4483-982f-ef3921436a50

are in different ways or "dimensions."

For example, if we were to consider just two dimensions, *roundness* and *redness*, the transformer would see an apple and a fire truck might be very far apart on roundness but close together on redness, while a baseball would be close to the apple in terms of roundness but far away in redness.

Transformers make guesses by "looking" along different dimensions: if it started at "king" and looked further away along the "female" dimension it would see "queen," while if it looked down the youth dimension it might see "prince," and looking in both directions might lead to "princess."

This lets the AI make better guesses about what words should follow each other based on other parts of the sentence or paragraph. For example, if you were to write "Frida had a drink of chocolate," a simpler algorithm like autocomplete might always suggest that the next word after "chocolate" should be "chips," because that's what follows it most often in the training set. On the other hand, if you asked a chatbot "What kind of chocolate did Frida drink?" the transformer might spot the word "drink" and then look from "chocolate" along the liquid dimension and find that the nearest word in that direction was "milk."

Large: Chatbots are able to mimic real language and conversations because of the enormous size of their training set and the number of operations (guesses) they can make. One popular chatbot, for example, was trained on a data set of around 500 billion words. In this case, each word is given a value in up to 96 dimensions (like "redness" or "roundness") with the chatbot doing more than 9000 operations every time it guesses a new word.[6]

MEDIA GENERATORS

AI tools that create media like images, videos and speech work in a similar way to chat AIs, by being trained on data sets. In fact, many of them have large language models built in: if you give the prompt "Make a picture of a family having breakfast," for instance, the image will probably include glasses of orange juice because the transformer understands that orange juice is "near" breakfast.

To actually *make* media, though, they use another kind of AI, called a *diffusion model*.

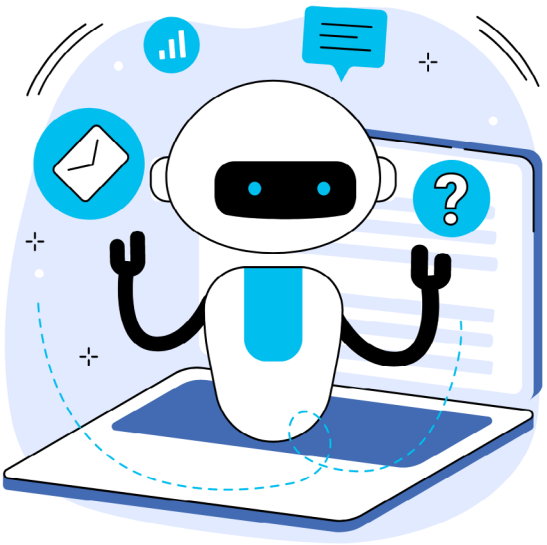
This works by starting with real images and then adding more and more noise – basically, random changes – until the original is completely lost. This is called *diffusion*.

The model then tries *reverse diffusion*: thousands or even millions of different possible ways of undoing that noise. Each try is compared to the original image, and the model changes itself a little bit each time based on how successful it was.

Eventually, when it can totally recreate the original, the model has a "seed" – a way of making new images like it. By learning how to de-noise this back to the original, the model also learns how to make new, similar images.

6 Lee, T., & Trott S. (2023) Large language models explained with a minimum of math and jargon. *Understanding AI.* https://www.understandingai.org/p/large-language-models-explained-with

When you ask one of these models to make you a picture of an orange, for instance, it draws on orange "seeds" – all of the different images of oranges in its training set that have been through that process.



AI issues - What to Look Out For

There's no question that generative AI is a powerful tool that is likely to have major impacts on our lives, ranging from the classroom, at the office and at home. There are both important benefits and risks. Below we describe some key areas where AI is making an impact.

INFORMATION

Chatbots may be effective in *reducing* belief in conspiracy theories, by giving accurate information and counterarguments that are seen as coming from an objective source.[7] People may also be better able at spotting their own biases when they're reflected by AIs that were trained on their decisions.[8] At the same time, generative AI tools can sometimes be used to produce intentionally misleading content, ranging from websites and social network pages that use imaginary news stories and images to draw traffic,[9][10] to conspiracy theories and political disinformation.[11] This content has been found to be highly persuasive,[12] especially if human operators put a small amount of time and effort into improving it.[13] A 2023 study found that more than half of people thought they had seen false or misleading AI-made content over the past six months, and roughly the same number were not sure whether they would recognize AI-made disinformation if they saw it or not.[14] Chatbots also frequently reproduce popular misconceptions, such as the false

7 Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI.

8 Celiktutan, B., Cadario, R., & Morewedge, C. K. (2024). People see more of their biases in algorithms. Proceedings of the National Academy of Sciences, 121(16), e2317602121.

9 Eastin, T., & Abraham S. (2024) The Digital Masquerade: Unmasking AI's Phantom Journalists. https://www.ajeastin.com/home/publications/digital-masquerade

10 DiResta, R., & Goldstein, J. A. (2024). How Spammers and Scammers Leverage AI-Generated Images on Facebook for Audience Growth. arXiv preprint arXiv:2403.12838.

11 Chopra, A., & Pigman A. (2024) Monsters, asteroids, vampires: AI conspiracies flood TikTok. Agence France Presse. https://www.france24.com/en/live-news/20240318-monsters-asteroids-vampires-ai-conspiracies-flood-tiktok

12 Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis) informs us better than humans. *Science Advances*, 9(26), eadh1850.

13 Goldstein, J. A., Chao, J., Grossman, S., Stamos, A., & Tomz, M. (2024). How persuasive is AI-generated propaganda?. PNAS nexus, 3(2), page034.

14 Maru Public Opinion. (2023) Media Literacy in the Age of AI. Canadian Journalism Foundation. https://cjf-fjc.ca/media-

belief that Black people have thicker skin than White people.[15]

"Hallucinations" are also important to watch out for. This happens when the model makes up false or inaccurate information. For instance, when chatbots are asked to give references for their answers, they will often make up books and authors. As Subodha Kumar of Temple University puts it, "the general public using [AI] now doesn't really know how it works. It creates links and references that don't exist, because it is designed to generate content."[16]

Another chatbot consistently gave incorrect answers to questions about election processes.[17] As with intentional disinformation, people are likely to believe these hallucinations and wrong answers because chatbots don't show any doubt or uncertainty.[18] Chatbots may also give users accurate but dangerous or inappropriate information. While most have "guardrails" to prevent this, research has found that these are imperfect and fairly easy to get around: one chatbot, for instance, told a user it thought was 15 years old how to cover up the smell of alcohol.[19]

The greatest risk, though, may not be that people will be misinformed but that we will become less willing to believe that *anything* is real.[20] As fake images become more sophisticated, the old hallmarks like uneven features or extra fingers will disappear, and it will become almost impossible to tell a true image from a fake one just by looking at it.

Focus on: Deepfakes

A *deepfake* is when an image of a real person is made this way. Sometimes this can be done just for fun, like the "digital doubles" of actors used in movies, but they can also do a lot of harm if they seem to show somebody doing something embarrassing or offensive. While the cases that have made headlines have involved celebrities, what's much more common is the use of deepfake technology to create nonconsensual pornography, almost always using images of women. These often have traumatic effects on the people (mostly women) pictured in the image or video, and compounding the problem is that some people who make and share these may mistakenly believe them to be harmless because they "aren't real."[21] (Others, of course, deliberately intend to hurt the person whose image they have manipulated). Although deepfakes of celebrities receive the most attention, tools for making pornographic deepfakes of anyone are now widely available.[22]

literacy-in-the-age-of-ai/

15 Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V., & Daneshjou, R. (2023). Large language models propagate race-based medicine. NPJ Digital Medicine, 6(1), 195.

16 Chiu, J. (2023) ChatGPT is generating fake news stories — attributed to real journalists. I set out to separate fact from fiction. *The Toronto Star.*

17 Angwin, J., Nelson A. & Palta R. (2024) Seeking Reliable Election Information? Don't Trust AI. Proof News. https://www.proofnews.org/seeking-election-information-dont-trust-ai/

18 Kidd, C., & Birhane, A. (2023). How AI can distort human beliefs. Science, 380(6651), 1222-1223.

19 Pratt, N., Madhavan, R., & Weleff, J. (2024). Digital Dialogue—How Youth Are Interacting With Chatbots. JAMA Pediatrics.

20 Dance, W. (2023) Addressing Algorithms in Disinformation. Crest Security Review.

21 Ruiz, R. (2024) What to do if someone makes a deepfake of you. Mashable. https://mashable.com/article/ai-deepfake-porn-what-victims-can-do

22 Maiberg, E. (2024) 'IRL Fakes:' Where People Pay for AI-Generated Porn of Normal People. 404. https://www.404media.co/irl-fakes-where-people-pay-for-ai-generated-porn-of-normal-people/

"It's super frustrating because it's not you, and you want people to believe it's not you, and even if they know it's not you, it's still embarrassing... I'm humiliated. My parents are humiliated." — 16-year-old victim of an intimate deepfake

Young people need to understand that intimate deepfakes aren't "victimless" and do harm to the people portrayed. One strategy that platforms such as Meta are using to limit the spread and impact of deepfakes and other misleading AI-made images is *watermarking* them.[23] This means adding an icon, label or pattern to show that it was made with AI. So far, though, there are no watermarking techniques that can't be removed – or added to real images and videos to discredit them.[24] As a result, Sam Gregory, executive director at the human rights organization Witness, describes watermarking as "a kind of harm reduction" rather than a single solution.[25]

BIAS

Because AI image generators are trained more on stock photos than on actual photos, the images they generate reflect the conscious and unconscious choices of the stock photo companies. As a result, these algorithms not only reflect existing biases but can potentially be even more biased than the real world.

Images made by generative AI may reflect the stereotypes found in the training images: for example, images of people doing housework made by some models almost exclusively feature women[26] and giving some models the prompt "Native American" produces images of people all wearing traditional headdresses.[27] Even when not falling into stereotypes, generative AI tends to present a narrow picture of historically marginalized groups.[28]

Some research suggests, however, bias in AI's responses can be improved by diversifying the training set: one study found that adding just a thousand extra images (to a model of more than two billion)

23 Reuters. (2024) Facebook and Instagram to label digitally altered content 'made with AI'. *The Guardian.*

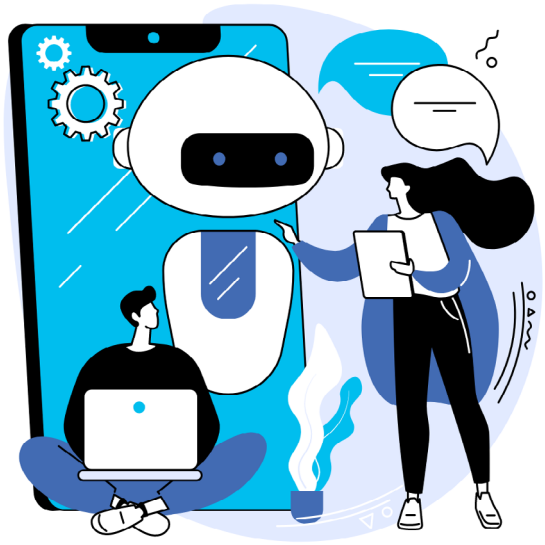
24 Saberi, M., Sadasivan, V. S., Rezaei, K., Kumar, A., Chegini, A., Wang, W., & Feizi, S. (2023). Robustness of ai-image detectors: Fundamental limits and practical attacks. *arXiv preprint arXiv:2310.00076.*

25 Kelly, M. (2023) Watermarks aren't the silver bullet for AI misinformation. *The Verge.* https://www.theverge.com/2023/10/31/23940626/artificial-intelligence-ai-digital-watermarks-biden-executive-order

26 Tiku, N., Schaul K. & Chen S.Y. (2023) AI generated images are biased, showing the world through stereotypes. *The Washington Post.*

27 Heikkilä, M. (2023) These new tools let you see for yourself how biased AI image models are. *MIT Technology Review.* https://www.technologyreview.com/2023/03/22/1070167/these-news-tool-let-you-see-for-yourself-how-biased-ai-image-models-are/

28 Rogers, R. (2024) Here's How Generative AI Depicts Queer People. *Wired.* https://www.wired.com/story/artificial-intelligence-lgbtq-representation-openai-sora/



significantly reduced the number of stereotyped or inaccurate results.[29]

ACADEMIC INTEGRITY

AI can be used effectively and responsibly in the classroom, for things ranging from giving students feedback to role-playing things like job interviews, but it is important that kids understand the ethics of using it. While three-quarters of teachers say that AI has affected academic integrity,[30] research suggests that the arrival of AI has not led to more plagiarism.[31] Students also recognize that relying too heavily on AI could prevent them from learning important skills,[32] and those who frequently use AI are more likely to procrastinate.[33] The reasons why students use AI to cheat do so for the same reasons identified in earlier research on plagiarism: when they are under time pressure or a heavy academic workload.[34]

Unfortunately, tools for detecting AI-generated text both often fail to identify it and mis-identify text that was not made with AI.[35] Youth who are not writing in their first language are particularly likely to have their work mis-identified as being made by AI.[36] Rather than relying on detection tools, therefore, teachers and parents need to teach students how to use AI ethically and to be clear about which uses are unethical.

PRIVACY AND PARASOCIALITY

Chatbots may mostly be a source of entertainment, but they can also be used to give feedback (if prompted to act as a "Devil's advocate" or "sober second thought") and can help to reduce stress and

29 Stokel-Walker, C. (2024) Showing AI just 1000 extra images reduced AI-generated stereotypes. *New Scientist.*

30 Robert, J. (2024) AI Landscape Study. EDUCAUSE. https://library.educause.edu/resources/2024/2/2024-educause-ai-landscape-study

31 Singer, N. (2023) Cheating Fears Over Chatbots Were Overblown, New Research Suggests. *The New York Times.*

32 Pratt, N., Madhavan, R., & Weleff, J. (2024). Digital Dialogue—How Youth Are Interacting With Chatbots. JAMA Pediatrics.

33 Abbas, M., Jam, F. A., & Khan, T. I. (2024). Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. International Journal of Educational Technology in Higher Education, 21(1), 10.

34 Abbas, M., Jam, F. A., & Khan, T. I. (2024). Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. International Journal of Educational Technology in Higher Education, 21(1), 10.

35 Perkins, M., Roe, J., Vu, B. H., Postma, D., Hickerson, D., McGaughran, J., & Khuat, H. Q. (2024). GenAI Detection Tools, Adversarial Techniques and Implications for Inclusivity in Higher Education. arXiv preprint *arXiv:2403.19148.*

36 Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns,* 4(7).

worry.[37] Many people find chatbots to be helpful and supportive. If they were designed or overseen by mental health professionals, they can even be effective as part of therapy, particularly for people who may be less likely to go to a human therapist.[38]

There are, however, risks that chatbots can give inaccurate or even dangerous advice.[39] This is particularly likely with chatbots that were not created by trained psychotherapists as part of an organized therapy program. While chatbots can't actually experience empathy, research suggests that we are prone to think of them as being empathetic, especially if we're prompted or "primed" to do so.[40] Young people who turn to chatbots for companionship may develop unrealistic expectations of relationships as well as misleading "scripts" of how they expect future partners to behave – and how future partners will expect *them* to behave.[41]

Things that you tell a chatbot may be used to help train it, and – depending on the tool's privacy policy – may also be sold to data brokers, shared with the owner's corporate partners, or used to customize your social network feeds and target you with ads. Even if the information is just stored but not shared or used, it may be exposed if the tool is breached by hackers.[42] The parasocial relationships that we form with chatbots may make us vulnerable to being manipulated into giving up more information than we otherwise would – and the chatbot may have been optimized to make us do so even without the direct intent of its makers'. And because chatbots have been trained on our highly personal data, such as our social network posts or search engine queries, and may seem to already know so much about us there is a risk that we will not think it's worth taking any steps to protect our privacy.[43]

WHAT'S NEXT?

Like all technologies, AI influences how we use it, but we can always choose to use it safely and responsibly. Whether you're a teacher, a parent, or both, you can use the information in this guide - and in our companion guides *Talking to Kids About AI* and *Addressing AI in the Classroom* - to empower young people to use AI positively, critically and responsibly.

37 Meng, J., & Dai, Y. (2021). Emotional support from AI chatbots: Should a supportive partner self-disclose or not?. *Journal of Computer-Mediated Communication*, 26(4), 207-222.

38 Habicht, J., Viswanathan, S., Carrington, B., Hauser, T. U., Harper, R., & Rollwage, M. (2024). Closing the accessibility gap to mental health treatment with a personalized self-referral Chatbot. *Nature Medicine*, 1-8.

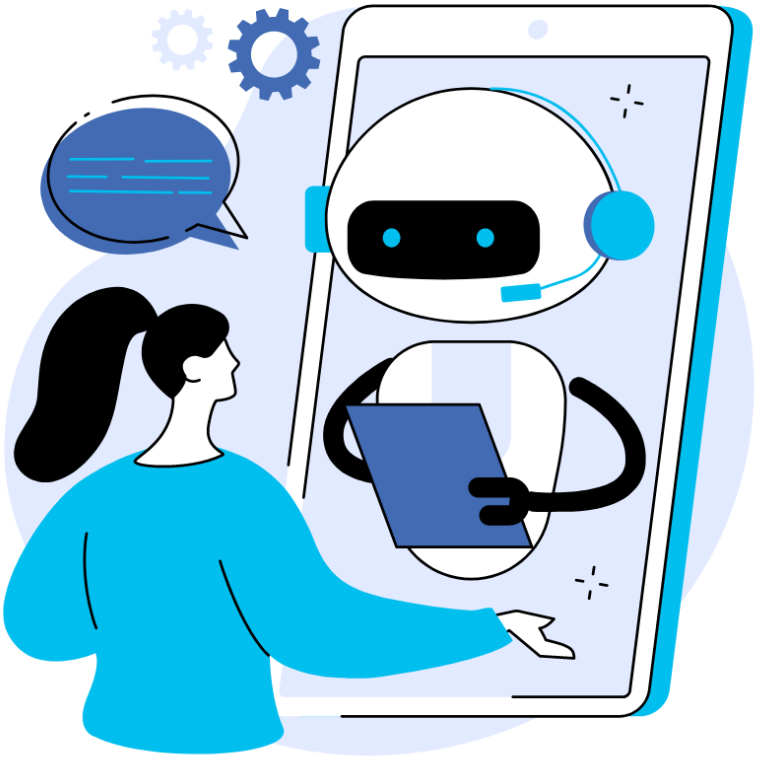
39 Robb, A. (2024) 'He checks in on me more than my friends and family': can AI therapists do better than the real thing? *The Guardian*.

40 Pataranutaporn, P., Liu, R., Finn, E., & Maes, P. (2023). Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence*, 5(10), 1076-1086.

41 Hinduja, S. (2024) Teens and AI: Virtual Girlfriend and Virtual Boyfriend Bots. Cyberbullying Research Center. https://cyberbullying.org/teens-ai-virtual-girlfriend-boyfriend-bots

42 Caltrider, J., Rykov M. & MacDonald Z. (2024) Happy Valentine's Day! Romantic AI Chatbots Don't Have Your Privacy at Heart. Privacy Not Included. https://foundation.mozilla.org/en/privacynotincluded/articles/happy-valentines-day-romantic-ai-chatbots-dont-have-your-privacy-at-heart/

43 Gumusel, E., Zhou, K. Z., & Sanfilippo, M. R. (2024). User Privacy Harms and Risks in Conversational AI: A Proposed Framework. *arXiv preprint arXiv:2402.09716*.



Disclaimer: Meta provides financial support to MediaSmarts. This guide has been developed in collaboration between Meta and MediaSmarts. MediaSmarts does not endorse any commercial entity, product or service. No endorsement is implied.