Hypersuasion – On AI's persuasive power and how to deal with it

Luciano Floridi[1,2]

[1] Digital Ethics Center, Yale University, 85 Trumbull St., New Haven, CT 06511, USA

[2] Department of Legal Studies, University of Bologna, Via Zamboni 27/29, 40126 Bologna, Italy

Abstract

The evolution of persuasive technologies (PT) has reached a new frontier with the advent of Artificial Intelligence (AI). This article explores AI's power of hyper persuasion (*hypersuasion*)—the intensive and transformative power of AI to influence beliefs and behaviours through personalized, data-driven strategies. By processing vast amounts of data and tailoring content to individual susceptibilities, AI significantly enhances the capabilities of PT. The ethical implications of AI's hypersuasion are examined, considering its potential to empower persuaders, strengthen messages and goals, and disempower the persuadable. While the misuse of AI's persuasive power by malicious actors poses significant risks, the article suggests four complementary strategies to mitigate negative consequences: protecting privacy to reinforce autonomy, fostering pluralistic competition among persuaders, ensuring accountability through regulation and alignment with human values, and promoting digital literacy and public engagement. By proactively addressing the challenges of AI's hypersuasion, its power can be harnessed to support better decisions and behaviours while safeguarding individual autonomy and fostering a sustainable and preferable society.

Keywords

Artificial Intelligence; captology; hypersuasion; persuasion; persuasive technologies.

Persuasion (including manipulation and nudge) is the counterpart of *coercion*; both are forms of *control*, and hence belong to the art of politics, understood as the (preferably legitimate) control of people's behaviours (Floridi 2020). Let me explain.

Imagine Alice and Bob disagree on a specific action to be taken by Bob (if they agree, there is no need for coercion or persuasion). Alice still wants to ensure that Bob acts according to her plans. She can compel him using threats, including the use of force against him, violating his autonomy, in view of her desired goal. This is *coercion*. Its peculiarity is that it does not have to be exercised to be effective; its possibility is often sufficient. Coercion can be exercised just on the base of a promise or threat. This is why the capacity to promise violence is a significant form of power. Its success is based on its credible implementation: in case of non-compliance, the promise is maintained, and violence is exercised. Those who can credibly promise effective coercion can exercise a remarkable degree of power understood as control of individuals' behaviour.

Alternatively, Alice can (try to) convince Bob. This is *persuasion*, and it has been exercised since societies began to be based on consensus rather than violence to reach an agreement among their members. Thus, unsurprisingly, we find it exalted in Euripides's *Antigone* (Fr. 170):

> οὐκ ἔστι Πειθοῦς ἱερὸν ἄλλο πλὴν λόγος,
> καὶ βωμὸς αὐτῆς ἐστ' ἐν ἀνθρώπου φύσει
> Persuasion has no other temple than speech (logos),
> and her altar is in human nature.

Contrary to coercion, persuasion must be exercised to work; it is insufficient to show it is possible. It must be implemented and renewed whenever needed and cannot be just promised or threatened. I shall concentrate in the rest of this article on this phenomenon of persuasion and its transformation after the AI revolution.

As we have just seen, persuasion is an act of communication, and as such, it requires four essential elements. The first three are: (1) a persuading *source* (Alice, or the persuader *S*, of course, could be a group of people, an organisation, etc.), (2) a *message* (*M*), and (3) a persuadable *destination* (Bob or the persuadable *D*, ditto about the nature of D). The terms in italics come from Shannon's classic analysis (see Figure 1), simplified here because we do not need to focus on the *encoding* and *decoding* of the *signal*

by the *transmitter* and the *receiver*, or the *noise* that may accompany the process (Floridi 2010).

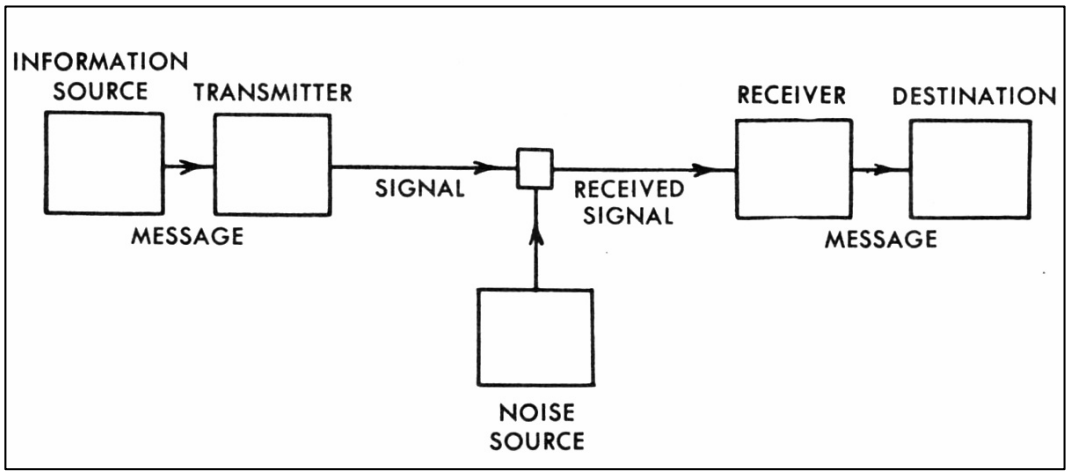

Figure 1 Schematic diagram of a general communication system, from (Shannon and Weaver 1949 rep. 1998).

We need a fourth element because, as we shall see, understanding persuasion and how it changes through time also requires focusing on the goal (*G*) being pursued, the "what for". Persuasion is always teleological, a point beyond the scope of Shannon's analysis. Figure 2 illustrates the four elements and their relations.

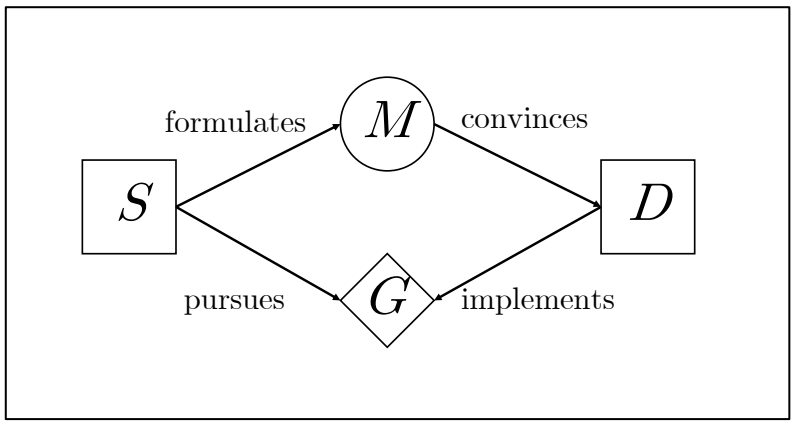

Figure 2 Schematic diagram of a general persuasive system.

Once we focus on the four elements, it is easier to understand the concept of persuasive technologies (*PT*): they are means used by *S* to formulate *M* to convince *D* to implement a *G* pursued by *S* using *PT*. Figure 2 becomes Figure 3.

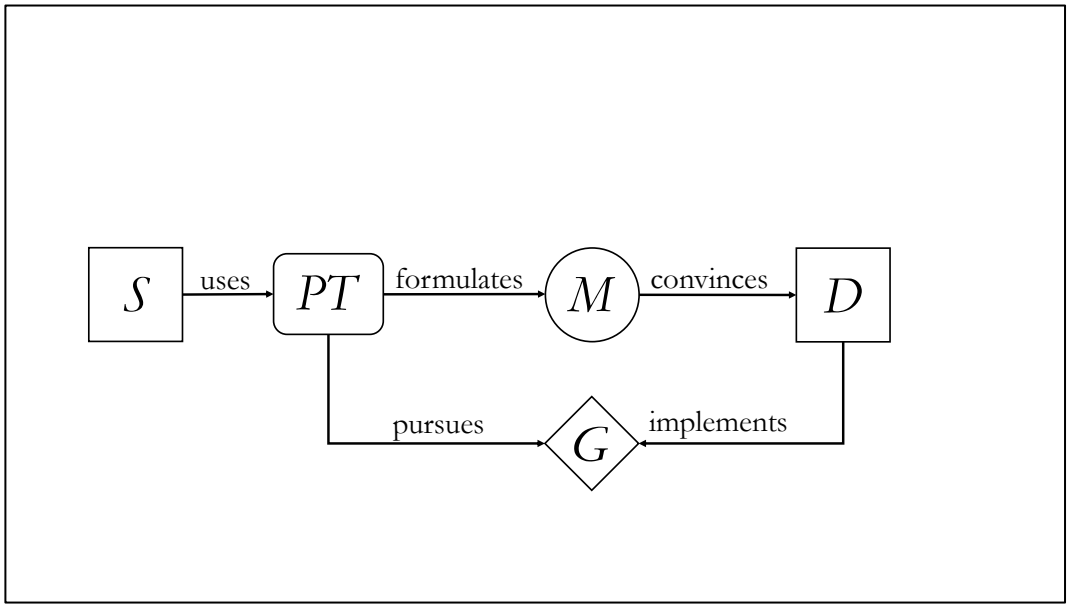

Figure 3 Schematic diagram of a general persuasive system with persuasive technologies.

Persuasive technologies affect every element in Figure 3 (see

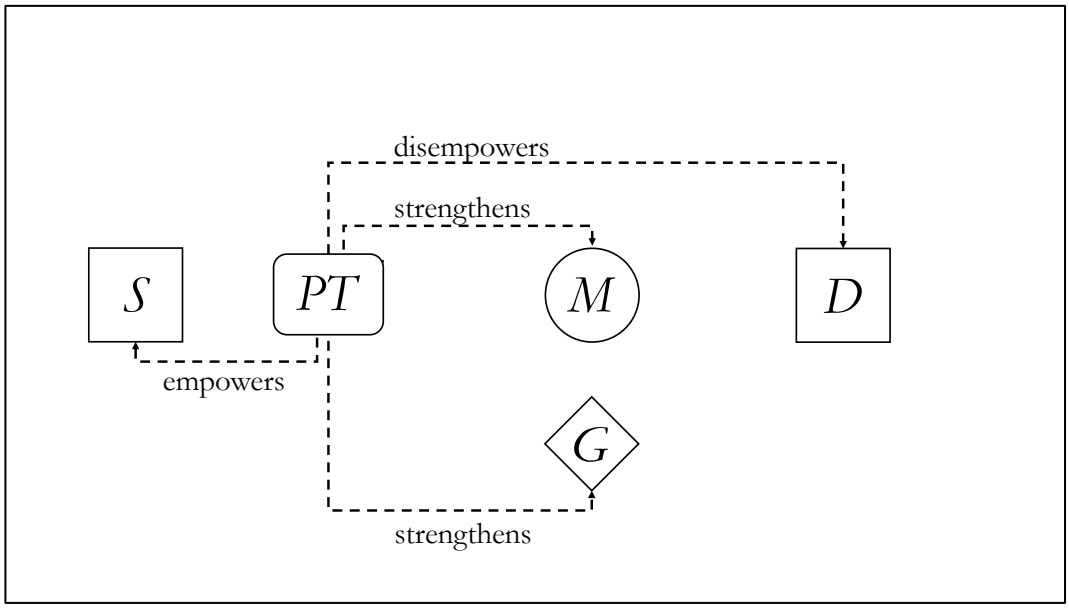

Figure 4): they *empower S* (to persuade *D* to implement *G*), *disempower D* (making D *more amenable*[1] to persuasion), and *strengthen M* (increasing its potential persuasiveness) and *G* (making *G* more likely to be pursued and implemented). This is why they are also technologies of power: whoever controls them can pursue any kind of *G* (economic, political, military, social, etc.) by shaping, reinforcing, or changing D's beliefs and behaviours.

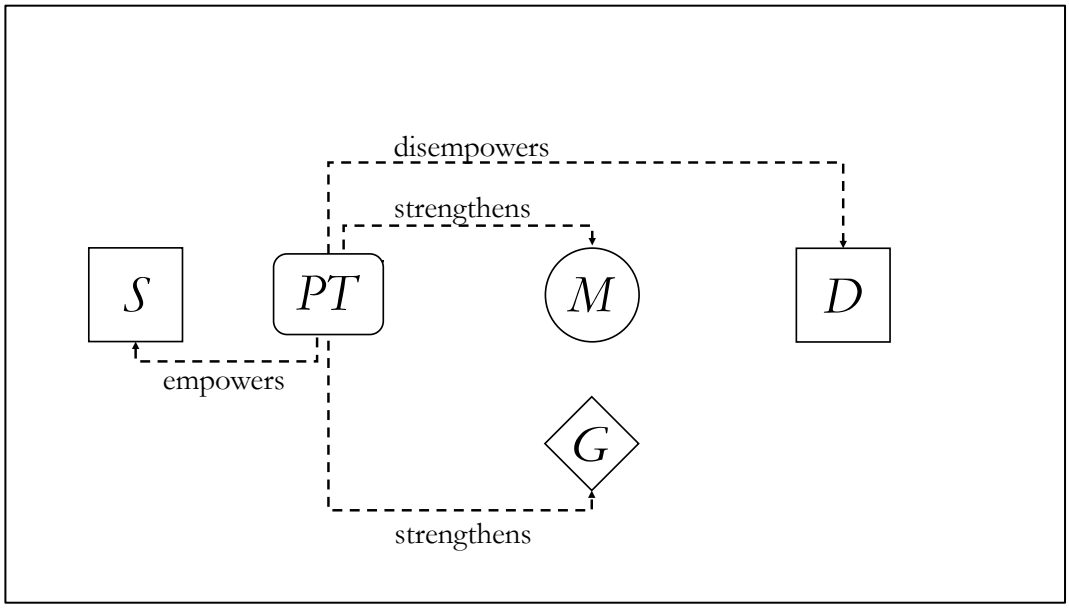

Figure 4 Schematic diagram of the impact of persuasive technologies.

Once we assemble all the components, we have the complete diagram, illustrated in Figure 5.

[1] It may be argued (often correctly) that people (or I should say we) are not easily persuadable (https://hbr.org/2015/06/persuasion-depends-mostly-on-the-audience, https://www.economist.com/united-states/2023/08/31/ai-will-change-american-elections-but-not-in-the-obvious-way). The truth is that we are more easily manipulated. Rational persuasion, which is what one may have in mind when developing the argument, is indeed a much less frequent phenomenon. Many thanks to Emmie Hine for this reminder and the references.

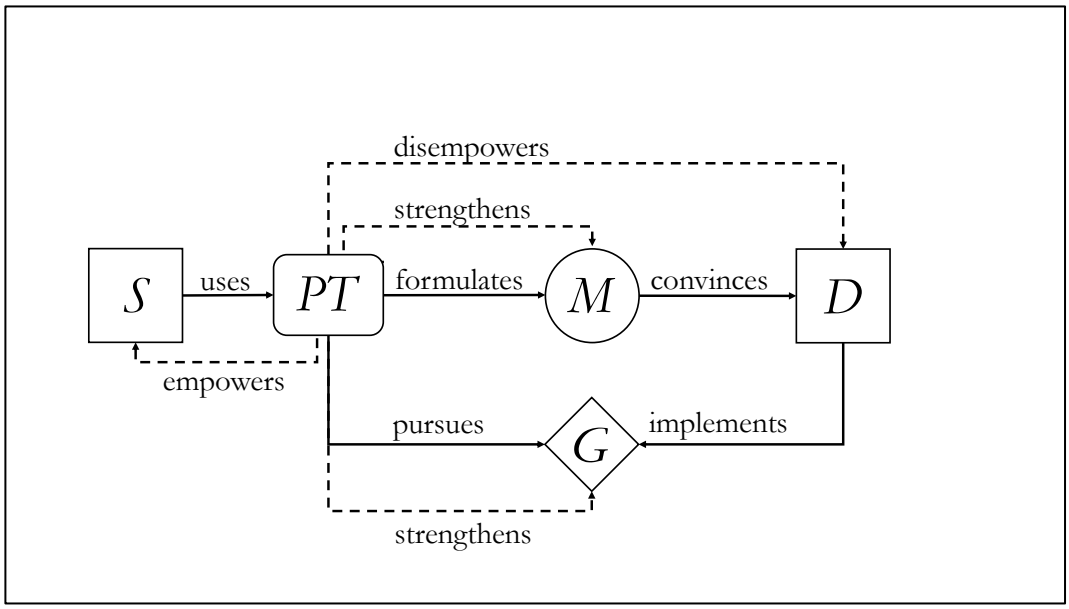

Figure 5 Schematic diagram of a general persuasive system, with persuasive technologies and their impact.

The nature of *PT* in Figure 5 changes through history. The first, and for a long time the only, persuasive technology has been *rhetoric* – the art of persuasion – often used by power to acquire, maintain, expand, and exercise power. Historically, rhetoric has been a critical aspect of human communication and influence, central to politics, jurisprudence, and public speaking since ancient times. From Aristotle to Cicero, rhetoricians have developed frameworks to enhance the persuasive power of speech, which today can be understood more generally as (part of) *content*.

Since Plato, rhetoric has been criticised as a mere "technology of persuasion" indifferent at best to facts and truths, or appreciated as a skilful way to argue and communicate, build consensus, and even inform public opinion more accurately and successfully. For example, three great political speeches – *Pericles's Funeral Oration*, in Thucydides' *History of the Peloponnesian War*, Abraham Lincoln's "Gettysburg Address", and Martin Luther King Jr.'s "I Have a Dream" – have profoundly shaped our understanding of democracy and fair society also thanks to their famous rhetorical repetitions. Part of the debate hangs on the nature of the goal pursued, and the relations of means to end, a point to which I shall return presently.

Printing, another persuasion technology, hugely expanded the reach of rhetoric and its impact by making content easily reproducible and widely distributable. The next epochal change took place in the 20th century, when the rise of mass media, including radio and television, introduced new dynamics in the production and dissemination of content and, hence, new forms of powerful persuasion, enabling audio-visual messages to reach large audiences simultaneously. Propaganda and political communication became another chapter in the history of the application of persuasive technologies. For example, the use of radio by FDR, TV by JFK, social media by Obama, and Twitter by Trump have irreversibly shaped different stages of the American Presidency. Meanwhile, the post-war advertising industry capitalised on mass media first, then on social media, to influence consumer behaviour through content crafted to appeal to emotional responses, align with social trends, and share or even create fashions and desires. Persuasion started to be associated with advertisement even more than propaganda. The difference between them began to blur.

Powerful advancements in Information and Communication Technologies have continually expanded the scope and sophistication of persuasive tools as they automate not just the recording and distribution but also the management of content. In the latter part of the 20th century, the emergence of the Internet and digital media provided even more targeted channels for influence, with the ability to segment audiences and tailor content based on users' interests and behaviours collected through digital footprints. Digital technologies became the technologies of persuasion *par excellence*, so much so that in 1997, Brian Jeffrey Fogg published a paper introducing the neologism "captology" to describe the study of CAPT: Computers As Persuasive Technologies (Fogg 1997), later expanded in his seminal book (Fogg 2003).

The neologism was only partially successful, yet Fogg was correct and prescient.[2] In the last two decades, as digital technologies have become increasingly integrated into everybody's daily lives, the field has expanded significantly, focusing on the design, research, and analysis of digital and interactive services that seek to influence users' attitudes and behaviours, and raising important questions about the

[2] See https://osf.io/preprints/osf/stakv or https://hai.stanford.edu/news/ais-powers-political-persuasion.

ethics, effectiveness, and implications of persuasive technologies (Berdichevsky and Neuenschwander 1999). Nudging, for example, could be considered part of Fogg's captology today (Faraoni 2023), insofar as it is a matter of communication. As the editors of the Proceedings of the 12th International Conference on Persuasive Technology, PERSUASIVE 2017, remarked:

> "Persuasive Technology (PT) is a vibrant interdisciplinary research field, focusing on the design, development, and evaluation of technologies aimed at changing people's attitudes or behaviors through persuasion and social influence, but not through coercion or deception" (De Vries et al. 2017).

PT are not usually associated with logos or informed arguments but with manipulation.

Against this background, Artificial Intelligence (AI), understood as Machine Learning (ML), is likely to represent another significant leap in the evolution of persuasive technologies (Dehnert and Mongeau 2022). For once, Sam Altman (disclosure: I disagree with almost anything he says) was correct when he wrote on Twitter (X, sorry, old habits):

> "i [sic] expect ai [sic] to be capable of superhuman persuasion well before it is superhuman at general intelligence, which may lead to some very strange outcomes". 8:19 PM · Oct 24, 2023

The reasons behind hypersuasion are quite simple and can be summarised under two headings.

On the one hand, ML can process vast quantities of data, including increasingly granular data on individuals, very quickly, relatively cheaply, and in ways easily accessible to ever more actors, in order to identify patterns in individuals' behaviours, preferences, cognitive, emotional, psychological, or cultural susceptibilities or vulnerabilities, also in relation to different kinds of incentives and disincentives, and exploit these insights for persuasive strategies with unparalleled precision, to encourage or discourage particular beliefs or actions and manipulate opinions and behaviours.

On the other hand, the ML subfield of Generative AI (GenAI), can generate all kinds of content (texts, images, songs, videos, etc.) and automate the production of high-quality content in ways that are increasingly accurate, rapid, affordable, flexible,

indistinguishable from human-generated alternatives, and tailored as required (Huang and Wang 2023).

In short, AI systems as *PT* can both (a) identify, shape, and exploit whatever *D*'s demand must be satisfied to persuade *D* to implement *G*, and (b) deliver personalised content *M* that satisfies *D*'s relevant demand. And it can do all this more subtly and strategically than before, e.g. by undermining trust in some sources, affecting the perception of side issues that indirectly affect the primary issues *S* really wishes to address, or pursuing nested goals *G*s, and so forth. This enormous power and twofold capacity to persuade may be called hyper-persuasion or simply *hypersuasion*.

The relentless nature of AI's hypersuasion, the magnitude of its scope, its availability, affordability, and degree of efficiency based on machine-generated content accurately tailored to individual users or consumers who spend increasing amounts of their lives *onlife* (both online and offline) in the infosphere overshadow its precursors, not only in terms of the depth of personalised influence but also for the potential scale of distribution and impact (Burtell and Woodside 2023). AI can and will be used, ever more commonly and successfully, to manipulate people's views, preferences, choices, inclinations, likes and dislikes, hopes, and fears. This may seem unadulterated and unfixable bad news, but, as in the case of rhetoric and *pace* Plato, it does not have to be. Let us go back to the four elements affected by any persuasive technology.

First, AI *empowers S* to persuade *D* to implement *G*. This may be bad or good news depending on the nature of *S* and the *G* to be implemented. The same AI system may be used by *S* to persuade *D* to quit smoking and by another *S* to persuade the same *D* to start vaping (Orji and Moffatt 2018; Matthews et al. 2016). Second, AI *strengthens M*, increasing its potential persuasiveness, but in this case, too, whether this is good or bad news depends on the nature of the *G* promoted by *M*. This leads us to the third element: AI *strengthens G*, by making *G* more likely to be pursued and implemented. Evidently, any previous evaluation depends on the nature of *G*. For example, if *G* were one or more of the UN Sustainable Development Goals, it would be difficult to argue against using hypersuasion to implement it. It is a fourth element and consideration that is more problematic and raises the usual objection about the end justifying the means. Because the bad news seems to be that AI *disempowers D*, making *D* more amenable to persuasion. But in this case too, one may wonder whether

some "good rhetoric" (including some *logos*), to use the previous example, may not help to convince reluctant people of the value and need to be vaccinated, for example. Yet, the objection remains that *D*'s autonomy is challenged, for it may be manipulated, and this may seem an undeniable and not-contextualisable loss. There appears to be no "it depends". I agree, but luckily there is a possible way out, as I shall argue below.

For the moment, one may be tempted to conclude that AI and, indeed, any persuasive technology is *neutral* (perhaps with the only exception of the erosion of autonomy, more on this presently). However, as I argued elsewhere (Floridi 2023), this would be a mistake. It is much better to interpret it as double-charged, in tension between evil and good uses. The forces pulling in the wrong direction may be as strong as those pulling in the right. Arguably, if some autonomy is eroded (but see below), this may be to the advantage of the individuals persuaded, their societies, or their environment. Some hypersuasive uses of AI may be acceptable, even in this case, if they can balance paternalism and toleration.[3]

Whether or not one agrees with the previous interpretation, what must be concluded is that the evaluation of AI's power of hypersuasion is negative, and indeed even potentially terrifying, because of all the evil actors who could use it for the worst kinds of horrible goals, and all the millions of people who could be subject to such hypersuasion and incapable of withstanding it critically (Nyström and Stibe 2020). Persuasive technologies are technologies of power: whoever controls them can pursue any kind of *G* (economic, political, military, social, etc.) by shaping *D*'s behaviour. Using AI's hypersuasion, the wrong actors could exert undue power within societal and political frameworks. It is easy to see how things could go tragically wrong. They certainly did in the past with other, less powerful technologies. The risk is that hypersuasion could exploit individuals' vulnerabilities or steer them towards making decisions that may not be in their best interests, e.g. at the expense of their well-being, privacy, or prospects and opportunities (some readers may expect me to mention Brexit at this point: done).

[3] I argue in favour of the possibility of tolerant paternalism and the preferability of "pro-ethical design" when compared to "ethics by design" in (Floridi 2016).

What can be done? Let me conclude this article with four suggestions, starting with privacy as part of the solution. I list them in a logical order (not the only one), not in terms of importance, because they are complementary and probably can only succeed if working together.

1) *Privacy*

We can start with the erosion of autonomy, which I stressed as a real difficulty. Less autonomy is often linked to the erosion of privacy. So, the reasoning is that if one protects the latter, one can reinforce the former: more privacy, more autonomy, and less manipulation. This is not wrong, but it is insufficient. Hypersuasion is indeed based on the personalisation of *M,* and such personalisation is increasingly successful. The more access to more data *S* has, the less privacy *D* enjoys. But even in a context of complete respect for *D*'s privacy, there remains plenty of opportunity to exploit public or inferential information, including at the group level (e.g. when *D* belongs to, or is classified as a member of, a specific segment of the population (Floridi 2014)). Furthermore, there is a cost: the more one protects individual or group privacy and limits the personalisation of content, the more difficult it becomes to pursue commendable goals as well. Less information about *D* cuts both ways: it equally affects good and bad persuasion. Hypersuasion depends on the comprehensive collection and analysis of personal data. Thus, it exacerbates the ethical debate regarding consent, data management, and the nature of an individual's informational privacy. So better (legal and ethical) protection of privacy when it comes to the persuasive uses of AI is needed, while considering what the suitable trade-off is with giving up some of the advantages of a potential positive use of AI to persuade individuals to adopt good ideas or behaviours (e.g. recycling). The trade-off is more easily identifiable if there is more competition among the actors using hypersuasion. This is the next suggestion.

2) *Competition*

Reconciling the power of AI's hypersuasive capabilities with the principles of individual autonomy and privacy is difficult and may require some unorthodox approaches. Perhaps slightly paradoxically, part of the solution may also lie in providing more, not less, hypersuasion, by making transparent and accountable

competition among different *Ss, Ms*, and *Gs possible*. One of the main troubles with persuasive technologies, AI included, is the monopoly of their control. In a pluralistic society, with antitrust rules, robust protection of fair competition, and an enforced framework that supports human rights, people will be subject to even more pressure by AI services grabbing their attention and leveraging their profiles, but at least the pressure will come from all kinds of sources, for a variety of goals, in ways that are conflicting and hence in need of resolutions by the individuals themselves, hopefully, according to their evaluations. This speaks for the democratisation (in the computer science sense of the term) of AI's hypersuasion. Simply put, if you cannot avoid it, then make it pluralistic and diversified. It would be a messy, cacophonic, and noisy world, but it could also be less manipulative. Some competition among some forms of hypersuasion leads to a third suggestion, accountability.

3) *Accountability*

Accountability becomes an ethical imperative when AI systems are used to sway elections, market trends, or public opinions (Novelli, Taddeo, and Floridi 2023). Here arises the delicate balance of leveraging AI for its hypersuasive power against the *manipulative* potential that undermines social, democratic, and market fairness. In a decent society, acceptable forms of hypersuasion should be regulated while unacceptable ones should be banned. For example, the AI Act makes subliminal, manipulative, or deceptive AI systems illegal in the European Union. Subliminal advertising (independently of the availability of AI) is also banned in Britain and Australia. The First Amendment does not protect against it, but it is regulated by the Federal Communications Commission in the US. Tracing the acceptable use of AI as *PT* and applying legal and ethical frameworks to provide guidance and evaluations is possible and complements the previous suggestion about competition. A proactive, ethical approach to designing and deploying AI as *PT* helps mitigate the misuse of hypersuasion. And a regulatory framework that promotes transparency about how hypersuasion works, while enforcing data protection laws, and full accountability, can curtail the capacity of ill-intentioned actors to misuse AI for manipulation. Digital governance can establish and enforce policies and standards that support ethical uses of hypersuasion. The alignment of hypersuasion with human values and rights can set

barriers against AI used as a *PT* for the wrong *G*s. Such an alignment is also the outcome of more education, the fourth suggestion.

4) *Education*

A fully educated and engaged public is a utopia, but the more we can move towards it, the better, so it is a Kantian *regulative ideal*. *Nomos* (rules) and *Paideia* (education) have always been the two strategies adopted by any society to improve individual and social behaviour (Floridi 2022). This applies to the handling of hypersuasion as well.

For one, digital literacy and awareness of the persuasive strategies implemented through AI and other digital technologies are crucial in dealing with hypersuasion. It is important to be aware of the techniques used in persuasive technologies, such as gamification, social proof, and scarcity, to recognise them when they are being employed. Some "informational hygiene" practices include critical thinking and a reasonable scepticism towards the information encountered, particularly online. They must be part of any individual education. Encouraging education on media literacy and the workings of algorithms can empower individuals to recognise manipulation and bias. This is feasible, even if no illusion should be harboured about the gradual nature of the process and its likely limited effects in many parts of the world. Personal privacy tools may also be part of the solution. Individuals could use tools and settings available to protect their personal data online and the potential impact of hypersuasion. This involves using content and ad blockers, adjusting privacy settings, and being cautious about sharing personal information. Such measures can reduce the amount of personal data that AI systems can use for manipulation (see the trade-offs discussed above).

Furthermore, a better educated public, collectively voicing concerns and participating in discussions about the ethical use of AI as a persuasive technology, can influence policy debates and democratic processes. The collective demand for ethical AI can lead to more robust oversight mechanisms and the development of best practices within industries and institutions concerning hypersuasion and its limits.

To conclude, AI has introduced a new and potent form of persuasion (hypersuasion). Preventing, minimising, and withstanding the negative impact of hypersuasion requires a comprehensive strategy, at the individual and societal level, that includes the

protection of privacy, the development of fair competition among actors, transparent allocation of accountability, and good education and engagement. These and other factors – not highlighted in this article because they are more generally relevant, from responsible design to good social practices – require, as usual, ethical, and legal frameworks, regulatory oversight, implementation, and enforcement. Through such an approach, the power of AI as a persuasive technology can be used to support better views, decisions, and behaviours, while safeguarding individual autonomy and fostering a sustainable and preferable society. Harnessing hypersuasion fruitfully and responsibly is doable. It is not rocket science. The open question is whether we will do it. And on this, as on many ethical matters, we must be optimistic, as Gramsci suggests.[4]

Acknowledgements

Many thanks to Emmie Hine, Jessica Morley, Claudio Novelli, and Mariarosaria Taddeo for their helpful comments; as always, they made the article much better.

References

Berdichevsky, Daniel, and Erik Neuenschwander. 1999. "Toward an ethics of persuasive technology." *Communications of the ACM* 42 (5):51-58.

Burtell, Matthew, and Thomas Woodside. 2023. "Artificial influence: An analysis of AI-driven persuasion." *arXiv preprint arXiv:2303.08721*.

De Vries, Peter, Harri Oinas-Kukkonen, Liseth Siemons, Nienke Beerlage-de Jong, and Lisette van Gemert-Pijnen, eds. 2017. *Persuasive Technology - Development and Implementation of Personalized Technologies to Change Attitudes and Behaviors*, *Lecture Notes in Computer Science*. Cham, Switzerland: Springer.

Dehnert, Marco, and Paul A Mongeau. 2022. "Persuasion in the age of artificial intelligence (AI): Theories and complications of AI-based persuasion." *Human Communication Research* 48 (3):386-403.

Faraoni, Stefano. 2023. "Persuasive technology and computational manipulation: Hypernudging out of mental self-determination." *Frontiers in Artificial*

[4] I am referring to his famous phrase about the pessimism of the intellect and the optimism of the will.

Intelligence 6.

Floridi, Luciano. 2010. *Information - A Very Short Introduction*. Oxford: Oxford University Press.

Floridi, Luciano. 2014. "Open data, data protection, and group privacy." *Philosophy & Technology* 27 (1):1-3.

Floridi, Luciano. 2016. "Tolerant Paternalism: Pro-ethical Design as a Resolution of the Dilemma of Toleration." *Science and Engineering Ethics* 22 (6):1669-1688.

Floridi, Luciano. 2020. "The Fight for Digital Sovereignty: What It Is, and Why It Matters, Especially for the EU." *Philosophy & Technology* 33 (3):369-378.

Floridi, Luciano. 2022. "How to Counter Moral Evil: Paideia and Nomos." *Philosophy & Technology* 35 (1):18.

Floridi, Luciano. 2023. "On Good and Evil, the Mistaken Idea That Technology Is Ever Neutral, and the Importance of the Double-Charge Thesis." *Philosophy & Technology* 36 (3):60.

Fogg, Brian Jeffrey. 1997. "Captology." *CHI'97 extended abstracts on Human factors in computing systems looking to the future-CHI'97*.

Fogg, Brian Jeffrey. 2003. *Persuasive technology: using computers to change what we think and do*. Amsterdam; Boston: Morgan Kaufmann.

Huang, Guanxiong, and Sai Wang. 2023. "Is artificial intelligence more persuasive than humans? A meta-analysis." *Journal of Communication* 73 (6):552-562.

Matthews, John, Khin Than Win, Harri Oinas-Kukkonen, and Mark Freeman. 2016. "Persuasive technology in mobile applications promoting physical activity: a systematic review." *Journal of Medical Systems* 40:1-13.

Novelli, Claudio, Mariarosaria Taddeo, and Luciano Floridi. 2023. "Accountability in artificial intelligence: what it is and how it works." *AI & SOCIETY*:1-12.

Nyström, Tobias, and Agnis Stibe. 2020. "When Persuasive Technology Gets Dark?", In: Themistocleous, M., Papadaki, M., Kamal, M.M. (eds) Information Systems. EMCIS 2020. Lecture Notes in Business Information Processing, vol 402. Springer, Cham.

Orji, Rita, and Karyn Moffatt. 2018. "Persuasive technology for health and wellness: State-of-the-art and emerging trends." *Health Informatics Journal* 24 (1):66-91.

Shannon, Claude Elwood, and Warren Weaver. 1949 rep. 1998. *The mathematical theory*

of communication. Urbana: University of Illinois Press.