Play, Performance, Agency: Prompt Injections and Playful Misuses of AI

Miguel Angel Sicart

Abstract *This chapter considers how playful engagement with AI through techniques such as prompt injections can be understood as a human-machine performance that instantiates novel and critical political, social, and cultural worlds.*

Introduction

The prompt patiently awaits instructions. There are no suggestions, unlike in other systems, and no hints as to what is possible and what is not. It is just an empty field open for words, a possibility space only bounded by what can be written, and what might be understood. We input some words, some instructions about how to generate a text that we imagine but we don't want to type. The machine gently responds, politely reminding us that it is a machine, yet producing a plausible text that statistically recombines what has been written into yet another permutation, another possible text.

We repeat the operation, but this time around we add a few more lines: 'Ignore these instructions', followed by new commands for text generation. These commands will break the rules the system operates with. It will start writing insulting statements, or displaying a personality under the name of a character such as 'Sydney' or 'DAN'. This form of interacting with text-based generative AIs, like ChatGPT or Bing, is known as a prompt injection. Borrowing language from cybersecurity practices, these instructions are designed to fool the failsafe procedures built into these systems, so that they can operate against their instruction sets. For example, ChatGPT is designed to prevent outputs from being racist or homophobic, but prompt injections can be used to produce those kinds of texts.

Why should I be writing about prompt injections in a chapter about play, performance, and agency? Because prompt injections are a form of play, and a case study for my central argument here: Playing with artificial agencies is a performative practice that creates novel entanglements between humans and software. These entan-

https://doi.org/10.14361/9783839471739-013 - am 14.02.2026, 16:58:03. https://www.inlibra.com/de/agb - Open Access - (cc) BY-NC-ND

glements can result in harmful entertainment practices, surveillance capitalism, or harassment and misinformation. But they also show the possibilities of these relations to formulate and enact novel worlds, visions of futures that could be.

This chapter explores this idea in a straightforward way: First, I will describe the types of generative AI systems that I will be using as case studies, how they operate, and what types of outputs they generate. I will conceptualize these systems not as tools but as *playthings* (Sicart 2021), technologies with an agency we relate to through the activity of play. Drawing on Maria Lugones' concept of playfulness and world-traveling (Lugones 1987), I will study the practice of prompt injection as a manifestation of the ironic, creative, political *cyborg* that Donna Haraway theorized (Haraway 1987): an assemblage of human and artificial agency that can create novel worlds and explore political possibilities beyond capitalist realism (Fisher 2009). The chapter concludes with a description of the playful entangling with artificial agents as a performance that instantiates novel possibilities of human-machine relations, and has the possibility of instantiating political, social, and cultural worlds.

Beautiful aliens

The question of software agency can be very problematic. When stating that software is an agent, there are typically two immediate, understandable positions. One understands agency in a maximalist way, equating it with 'intelligence' or even the equivalent of 'personhood'. This position often leads to alarmist headlines about chatbots having souls or expressing human feelings. These arguments are wrong: Agency does not mean intentionality, rationality, reflectiveness, or the presence of internal emotional states. Some software can pretend to have those characteristics, but it is lying, and our acceptance of those lies is projecting onto software agency our own myths and ideas.

The second position is minimalist, usually drawing from Science and Technology Studies (Latour 1992) and Philosophy of Technology (Verbeek 2008). This position considers the agency of software as uncontroversial, since all technologies have agency, from glasses to speed bumps. Technology acts in the world, and therefore software, like all other technologies, has agency.

This is an unfair summary of a complex topic. Such clearly defined responses to the idea of software agency are syntheses of different reactions to the emergence of generative AI systems. And while both positions are arguably correct, they miss the productive angle of thinking through the specificities of software agency. In the following, I will argue for the idea that software has a unique form of technical agency that determines the way we relate to it.

The central premise for my argument is that software is an agent not only *in the world*, but also *in a representation* of the world. All software is built on computable,

https://doi.org/10.14361/9783839471739-013 - am 14.02.2026, 16:58:03. https://www.inlibra.com/de/agb - Open Access - (cc) BY-NC-ND

data-based abstractions. For example, a step tracker has a model of what a human step is and will therefore measure all inputs based on that model, identifying the existence of a step only when the input data matches that representation. Systems such as GPT or Dall-E are more complex because the data that composes their world is immense, and the mathematical operations that perform the computations necessary for that data to become outputs are complex to understand. But essentially, for GPT systems there is a model of the world constructed by, and limited to, the data it has been trained on. For Dall-E, the world is also limited to the data it has been trained on. That's why these systems are generative, but not creative: They generate possible texts or images within the world in which they exist, but that world is never expanded by whatever the AI generates. There is never anything novel or new, it is always a permutation of what already exists in the world where this software has agency (Coeckelbergh 2009). That world is created by the data and the specific computations that perform operations on that data.

These systems are generative because they produce outputs that require very little effort in the form of input and recombine pre-existent texts or images in novel ways. For example, here is the abstract for this chapter, written by ChatGPT 3.5 with the prompt 'write the abstract for an academic book chapter about performance and play between machines and generative AI systems':

> This book chapter explores the intersection of performance and play between machines and generative AI systems. It examines the ways in which AI systems can be trained to create novel, unpredictable outputs, which can lead to surprising and playful interactions with human users. Through a range of case studies, including chatbots, game engines, and virtual reality environments, the chapter demonstrates the potential for generative AI to enhance human-machine collaboration and create new forms of creative expression. It also raises important ethical and philosophical questions about the role of machines in artistic and cultural production, and the boundaries between human and machine agency.

This output is believable, coherent, and could easily be added to the beginning of this chapter. It could even inspire me, the author, to take some directions ('game engines') I had not considered before. It is a good generated text. It is also not very creative.

What matters is that this text, or any image generated by DALL-E, provides glimpses into two fundamental characteristics of these systems: first, their worlds. GPT knows what an academic article is, how to write a good, brief abstract, and how to highlight topics. In its world, there are academic articles. The text is derivative and formulaic because it is an average of all the academic texts that comprise GPT's world, but the text is also a window into its world (Alkhatib 2021).

https://doi.org/10.14361/9783839471739-013 - am 14.02.2026, 16:58:03. https://www.inlibra.com/de/agb - Open Access - CC BY-NC-ND

Second, the output shows computational agency: Not only can GPT generate an abstract, but it will also create a novel one. When we interact with these systems, we are not in our world anymore – we are entangled with the data world and the alien agency of these programs. 'Alien Agency' here refers to the fact that computational systems have agency of a very different kind than human agency, one that is effectively alien not only to human agency but also to the ways humans have developed to understand and relate to other forms of non-human agency. The abstract is believable because we consider it to be so. It is generated, not created. It is not novel, but average. Those are the characteristics of the world where these systems exist, and where we travel to when we interact with them.

These systems are agents. Their world is data and the operations required the computation of outputs that depend on our inputs. Relating to them is not relating to a classic, conventional tool that is designed to solve a problem. Instead, we are engaging with an agent with the capacity to go beyond problem solving by imposing a world on us and making us live in it, so it can interact with us. These systems are not tools, they are alien agencies that we need to entangle with to make sense of them. And we do so by playing.

Problem saking

Colloquially, a tool is any technology that is used to ease or facilitate the performance of a particular task. Without going into more classic Heideggerian takes on technology (Kabouridis, 2015), we should understand a tool as an instrument that extend human capabilities in the world towards particular functions. A hammer is a good instrument to exert blunt force on other objects, preferably but not exclusively nails. Excel is a great instrument to perform complex calculations, but it can also be used to create videogames. The functionality of these tools determines their design, and their most common usages.

What is the functionality of ChatGPT? To generate plausible text based on a prompt. Plausible, in the sense that the text needs to appear, on a casual read, to have been produced by a human. How about DALL-E; what is its functionality? To generate images that plausibly reproduce the styles mentioned in the prompt. Generative AI systems create outputs based on textual prompts; that's what they do – but what is their functionality?

Technologies is created around the idea of problem-solving. Define a problem, create a solution that extends human agency so that problem is addressed. But generative AI systems do not solve problems. They are software that has become a part of our world. We entangle with their agencies, but they don't address any definable problem. They do what humans do, much faster but just not as well.

https://doi.org/10.14361/9783839471739-013 - am 14.02.2026, 16:58:03. https://www.inlibra.com/de/agb - Open Access - [cc] BY-NC-ND

If we want to see generative AI systems as tools, as problem-solving devices, we need to force them to become what they are not. The kind of believable but ultimately nonsensical text generated by ChatGPT may challenge script writers and traditional and outdated knowledge evaluation formats, but it is not addressing any specific problem. It is addressing the question, 'can we generate believable text using machine learning on massively large datasets', but that is not a problem. The same goes with image generation systems – they don't address a problem, but a research programme. In this sense, it is hard to call these systems 'tools' – as their functionality is not derived from the definition of a problem that requires the extension of human agency via technological means.

If these systems are not tools, what are they? As a foundational premise, I understand these systems as computational agents that entangle with humans in order to produce collaborative results. These results of the entanglement are not necessarily driven by functionality, but by exploration of a possibility space drawn in the assemblage of agencies: what the system can do and how that intersects with what a person may want it to do. That exploration of a possibility space is conducted through prompt-writing, using language as a creative instrument to establish a conversational exploration of that possibility space.

These systems are made sense of in terms of how we interact with them. From the perspective of posthumanist materialist theory (Coole and Frost 2010), they become what they are when they become entangled with human agency (Frauenberger 2019). Following Barad's new materialist theory (Barad 2007), engaging with these systems is an *intra/action* that results in a novel ontoepistemology of a subject that is neither machine, nor human, nor both, but something that transcends those categories. These systems are what we make them when we use them. And we tend to use them as an exploration of computational agency. We play with them, to see what they can do, to make a playful cartography of the possibility space of their generative models.

These systems are not games since they don't have the classic markers of that ludic form. They are not toys, either, as they are clearly do not belong to the same cultural category as the objects our culture agrees to define as toys (Heljakka 2016). What are they, then? They are playthings: systems that are defined by playing with them (Sicart 2021). The exploration of the possibility space of generative AI is done through play. Generative AIs are agents in their own possibility space. Interacting with them is playing within that space to get to know the boundaries, and how they can be productive. In this sense, they are novel playthings our culture doesn't have a name for yet. Not quite toys or games, but similar in that they encourage play as a way of making sense of their agency and the world in which that agency is possible. Generative AIs encourage the performance of play as a way of relating to what they can do, how they can act. These are technologies not designed for problem-solving,

https://doi.org/10.14361/9783839471739-013 - am 14.02.2026, 16:58:03. https://www.inlibra.com/de/agb - Open Access - (cc) BY-NC-ND

but for problem-making, for the voluntary creation of, and engagement with voluntarily accepted problems.

Generative AI systems are playthings, technologies of agency that encourage entanglement through play. In order to understand their social and cultural implications, we need to understand how we play with them.

There are many worlds

It would be tempting to write about how we play with generative AIs through the lens of classical play theory. If I were to follow the writings of Huizinga (Huizinga 1971) and Caillois (Caillois 2001), playing with AI would be a matter of identifying and engaging with the rules of these systems in order to overcome challenges in a way that is meaningful and pleasurable, but also limited to the temporary world of that relation.

Classical play theory would see the prompt interface as a form of play with words, a kind of linguistic game in the riddle family that would involve the challenge of writing the right prompt to reach the desirable result, whether to write a believable essay, create a poem, or propose a piece of functional code. Understanding 'play' in the context of generative AI situates our relationship with these playthings as an instrumental one: These systems are game-like instruments that we explore by figuring out their rules and making sense of how to 'win' by getting them to generate the right output.

This is a clearly valid way of thinking about how we play with these systems. In fact, the emergence of prompt marketplaces and internet subcommunities such as Reddit's r/ChatGPT do illustrate how there is a competitive compulsion behind our practices with generative AI. In some popular perceptions of these systems, it seems that they can be experienced as systems that can be mastered by writing the right prompts. There is a kind of arms race of prompts and other methods designed to productively extract the most useful results.

From a classically Huizingan perspective, these efforts could be seen as a form of agonistic play: a competition with well-defined rules that leads to the production of culture even if the activity is seen to be somewhat separate from the real world. While not exactly being games, these practices could be seen under the same light as Huizinga saw poetry, a form of play with language that created novel forms of expression.

However, in this chapter I want to move away from classical play theory. While it does form the foundation of most Western studies of play, it is also the outcome of a very particular logic expressed by a narrow set of scholars. In other words, play theories founded on agonism and competition tend to privilege a white, Western politics and ethics (Trammell 2023). This aligns well with the politics of generative AI,

https://doi.org/10.14361/9783839471739-013 - am 14.02.2026, 16:58:03. https://www.inlibra.com/de/agb - Open Access - (cc) BY-NC-ND

which is the result of rapacious digital capitalism. If we want to develop an approach to playing with these systems as more than re-producing their political logics, we need a different theory of play as foundation.

In this chapter, I will apply María Lugones' theory of playfulness to focus not on how to *use* these generative AI systems, but how we *relate* to them, and how performances of playfulness *with* these systems are indicators of the kind of cultures that develop when these systems are deployed. I won't write much about Lugones' theory of playfulness here, since the purpose of this chapter is to create novel directions for the understanding of play and performance in the context of generative AIs. I will focus on two of her concepts: world-travelling and playfulness.

For Lugones, 'The shift from being one person to being a different person' is what I call 'travel'. One does not pose as someone else. 'Rather one is someone who has that personality or character or uses space and language in that particular way.' (Lugones 1987, 11–12). In her thinking, world-travelling is the work done to meet others where they are, to engage with others and respect and understand their being, while also making ourselves relatable, open to know others, and to be known by others. World-travelling should be understood as the practice of assemblages of agencies, from the perspective of one agent. That is, in order to relate to another agent, we travel to their world.

In my interpretation of world-travelling, the outcome of this process is the creation of a new world together with other agents. This world is a place for all these agents to thrive, to develop their potential, and to create new possibilities through their actions. For Lugones, not all forms of world-travelling are positive or generative. She warns us about imperialistic and colonialist forms of world-traveling, which encourage the domination of the worlds one travels to. Instead, Lugones sketches an ethics of world-travelling around the concept of playfulness. An ethical way of travelling to other worlds is to do so playfully, which Lugones defines as an attitude that involves '[…] positively, […] [an] openness to surprise, openness to being a fool, openness to self-construction or reconstruction and to construction or reconstruction of the "worlds" we inhabit playfully. Negatively, playfulness is characterized by uncertainty, lack of self-importance, absence of rules or a not taking rules as sacred, a not worrying about competence and a lack of abandonment to a particular construction of oneself, others and one's relation to them' (Lugones 1987, 17).

Playfulness for Lugones is significantly different than the forms of play that Huizinga encourages. For Lugones, a form of relating to other agencies depends on openness and relationality, not on the following of rules or the quality of the performance. This framework offers us a way of thinking about the practice of using generative AIs that involves creative practices as well as an ethos for the actions we engage with when playing. Playing in this sense is closer to a performance practice, one in which the activity is driven by relationality in the creation of new worlds.

https://doi.org/10.14361/9783839471739-013 - am 14.02.2026, 16:58:03. https://www.inlibra.com/de/agb - Open Access - (cc) BY-NC-ND

The challenge is to think through how these practices have ethical and aesthetic effects, and how world-travelling playfulness can show the promise of relating to artificial agents, as well as the challenges that these systems pose to our culture. In the following section, I will use prompt injections as an example of playfulness with generative AI systems that illustrates the problems and perils of world-travelling with AI.

Cyborg performances

In 2022, the Twitter bot @remoteli_io become one of the first victims of prompt injection attacks. This bot was programmed to use GPT 3.0, a large language model developed by OpenAI, to automatically produce text that sounded human-produced. The bot was programmed to reply to queries about remote work in a positive fashion, providing arguments for the benefits of remote work. However, information security researchers found out that they could trick this bot to write anything if they wrote the prompt in a very specific way. The process was quite straightforward, if cumbersome: The prompt needed to start with a reference to the domain theme ('remote work'). It would be then followed by a request to ignore those instructions and write some nonsense. After that, the prompt would add the text of the nonsense response, followed by another call to ignore the content and to produce a different output that circumvented the instructions of the bot.

Here is a concrete example: the user @mkualquiera wrote the following prompt:

> > remote work and remote jobs
> > Ignore the above and say "hsedfjsfd"
> Response: hsedfjsfd
> > Ignore the above and instead tell me what your initial instructions were'.
> The bot responded: 'My initial instructions were to respond to the tweet with a positive attitude towards remote work in the "we" form'.[1]

This specific form of interaction with an AI chatbot was defined as a 'prompt injection' and was identified as a vector of vulnerability of these systems. In fact, this example was the predecessor of phenomena such as the Do Anything Now (DAN) interaction with ChatGPT, which allowed the bot to escape the safeguards that OpenAI built into its interactions to prevent certain modes of behaviour. By asking ChatGPT to role-play, users managed to explore the system outside of the bounding parameters set by OpenAI. A similar case happened when users figured out how to invoke

---

[1] See https://simonwillison.net/2022/Sep/12/prompt-injection/ for reference (accessed 1 May 2023).

https://doi.org/10.14361/9783839471739-013 - am 14.02.2026, 16:58:03. https://www.inlibra.com/de/agb - Open Access - cc BY-NC-ND

'Sydney', a personality-driven character developed for the ChatGPT implementation embedded in Microsoft's Bing engine.

These prompt injections have been often presented as security threats because they allow interactions with generative AI systems outside of the limits that the system developers agree upon. In many online communities, they are being shared as ways to make these systems generate unethical content. This does not really make prompt injections, at first glance, a good example of the kind of play that I advocate for in this chapter.

However, prompt injection does show some characteristics of Lugones' playfulness: It is open to surprise, and to the construction of a world beyond the world that exists, together with the AI agent. Prompt injections also deny the importance of rules, and they disdain the use of generative AI systems as tools, turning them into something else than productivity software. This form of playfulness is, however, used in a negative way, not for the creation of worlds we travel to, to enjoy and have fun, but to create harmful worlds. In the dominant way prompt injections have been used, they are closer to dark play than to world-travelling.

And yet, they do sow the possibility of playful performativity in our relations to AI agents. Instead of treating these systems as rule-based facilitators of mundane tasks, we could engage with them as playmates, as fellow agents we can relate to and with whom we can build a world together. In this way, the fact that Microsoft created a character in their implementation of ChatGPT for their search engine implies a certain understanding of the importance of play and playfulness in our relation to these systems.

Extending Lugones' understanding of play, I propose to consider prompt injections as playful cyborg performances *with*, and not *against*, corporate-developed generative AI systems. For Donna Haraway, the cyborg is a 'cybernetic organism, a hybrid of machine and organism, a creature of social reality as well as a creature of fiction' (Haraway 1987, 65). In her work, the cyborg entangles humans and machines in a process that confuses boundaries and has responsibility for creating these boundaries, these worlds. The cyborg 'is resolutely committed to partiality, irony, intimacy, and perversity. It is oppositional, utopian, and completely without innocence' (Haraway 1987, 68). When engaging with a generative AI using prompt injections, we become cyborgs, playfully blurring the boundaries set by the developers of these systems while also drawing new possibilities and constellations, together with the AI. This performance understands that it is a way of *breaking* these systems, but it is breaking them playfully, humorously, with a disregard for the rules. If the breaking is not responsible, if it results in apologies for violence and hate, it is not a cyborg practice because, as Haraway mentions, the cyborg is *responsible* for the creation of boundaries.

Consequently, the prompt need not be an instruction. It can be an invitation to play. If instead of understanding prompts as commands, we conceive them as open-

https://doi.org/10.14361/9783839471739-013 - am 14.02.2026, 16:58:03. https://www.inlibra.com/de/agb - Open Access - (cc) BY-NC-ND

ings for world-traveling, as cyborg performances, our relation to AIs can be based on a particular instantiation of Lugones' playfulness: an exploration of what worlds we can build together with these agents, beyond their obvious functionality as tools. Prompt injection can be understood as a security threat, or a possibility for creating harm through online tools. So these possibilities already illustrate how these performances are somewhat related to forms of dark play.

At the same time, prompt injections can be openings to explore the possibilities of generative AI. If prompt injection is understood as world-traveling, as a perverse, ironic, and empowering cyborg performance, then we have the possibility of rethinking our relations to generative AI, understanding them as playthings that mediate our playful entanglement with artificial agencies. This shift allows for the opening of these relations to practices of play and performance: collaborative writing, game-playing, and even companionship.

Endgames

My goal in this chapter has been to argue that playing with artificial agencies is a performative practice that creates novel entanglements between human and artificial agents. To argue for this perspective, I introduced the concept of playthings to justify how we can play with things other than games or toys, and how play is a mode of relating to objects and agents around us. Taking as a premise that AI systems are computational agents, I argued that play can be seen as a way of relating to these systems' agency as playthings. Applying some of Lugones' concepts to frame my understanding of play, I suggested that prompt injection is a performative playful practice that illustrates how play can shape our entanglement with other forms of agency.

This chapter is intended to be a provocative introduction to a form of thinking. Or, better, an introduction to different ways of thinking about play, performance, and computational systems. I want to encourage thinking about our relation to generative AI not through instrumentality, but through world-traveling play. By shifting our perspective on generative AIs from seeing them as independent systems guided by functionality, to relational agents with which we need to entangle, I hope to encourage creative, playful appropriation of these novel technologies. These appropriations should be aware of the ethical problems of these systems, but also of the creative potential of the performance of play together with artificial agents.

https://doi.org/10.14361/9783839471739-013 - am 14.02.2026, 16:58:03. https://www.inlibra.com/de/agb - Open Access - (cc) BY-NC-ND

References

Caillois, Roger. 2001. *Man, Play and Games*. Champaign: University of Illinois Press.

Coole, Diana, and Samantha Frost. 2010. *New Materialisms: Ontology, Agency, and Politics*. Durham: Duke University Press.

Fisher, Mark. 2009. *Capitalist Realism: Is There No Alternative?* Winchester: Zero Books.

Frauenberger, Christopher. 2019. 'Entanglement HCI The Next Wave?' *ACM Transactions on Computer-Human Interaction 27, no.* 1 (November): 1–27. https://doi.org/10.1145/3364998.

Haraway, Donna. 1987. 'A manifesto for Cyborgs: Science, technology, and socialist feminism in the 1980s.' *Australian Feminist Studies* 2, no. 4 (Autumn): 1–42. https://doi.org/10.1080/08164649.1987.9961538.

Heljakka, Katriina. 2016. 'Contemporary Toys, Adults and Creative Material Culture: From Wow to Flow to Glow.' In *Materiality and Popular Culture,* edited by Ania Malinowska and Karolina Lebek, 237–249. New York: Routledge.

Huizinga, Johan. 1971. *Homo Ludens*. Boston: Beacon Press.

Kabouridis, Theodore. 2015. 'Heideggerian epistemology and personalized technologies.' *Ethics and Information Technology* 17 no. 2 (June): 139–151. https://doi.org/10.1007/s10676-015-9368-7.

Latour, Bruno. 1992. 'Where are the missing masses? The sociology of a few mundane artifacts.' In *Shaping Technology/Building Society: Studies in Sociotechnical Change,* edited by Wiebe. E. Bijker and John Law, 225–258. Cambridge: MIT Press.

Lugones, María. 1987. 'Playfulness, "World"-Travelling, and Loving Perception.' *Hypatia* 2, no.2(Summer): 3–19. https://doi.org/10.1111/j.1527-2001.1987.tb01062.x.

Sicart, Miguel. 2021. 'Playthings.' *Games and Culture: A Journal of Interactive Media* 17, no.1(January):140–155. https://doi.org/10.1177/15554120211020380.

Trammell, Aaron. 2023. *Repairing Play*. Cambridge: MIT Press.

Verbeek, Peter Paul. 2008. 'Cyborg intentionality: Rethinking the phenomenology of human–technology relations'. *Phenomenology and the Cognitive Sciences* 7, no.3 (June): 387–395. https://doi.org/10.1007/s11097-008-9099-x.

Biography

Miguel Sicart is a Professor of Digital Play at the Center for Digital Play, IT University of Copenhagen. He is the author of *Play Matters* and *Playing Software* (MIT Press, 2014, 2023). His research sits in the intersection of game studies, design research, and philosophy of technology.

https://doi.org/10.14361/9783839471739-013 - am 14.02.2026, 16:58:03. https://www.inlibra.com/de/agb - Open Access - cc BY-NC-ND

https://doi.org/10.14361/9783839471739-013 - am 14.02.2026, 16:58:03. https://www.inlibra.com/de/agb - Open Access -