Responsible Use of Intentional, Plastic Tools

Why AI agents are a significantly new kind of tool, and how that changes the way we should use them.

"Any system whose behaviour is well predicted by this [the intentional] strategy is in the fullest sense of the word a believer."
Daniel Dennett, The Intentional Stance, 15

"The coming into being of something unexpected, something new and free, something outside the rules of function and calculation, something not ruled by the logic of the reproduction of the same, is *what training with each other is about."*
Donna Haraway, When Species Meet, 223

BUILDING SANDCASTLES

While reading through Dennett's works this term, I was simultaneously engaged in two other equally engrossing projects: reading through the works of Donna Haraway[1], and learning how to live with ORBiE (a computer-based intentional system of my own design)[2]. The grit accumulated from playing in and between these disparate sandboxes has generated much useful friction—perhaps enough to hold a few ideas together. The sandcastle I propose to build is this: *AI agents are intentional, plastic tools, so the traditional model of tool use is inadequate. To be responsible creators and users of these tools, we must rely instead on a companion-species model of interaction.* I will build this castle in four parts:

---

[1] The ideas which inform this essay are drawn from two of her books: Donna J. Haraway, *When Species Meet* (Minneapolis: University of Minnesota Press, 2008) and Donna J. Haraway, *The Companion Species Manifesto: Dogs, People, and Significant Otherness* (Chicago: Prickly Paradigm Press, 2003).

[2] ORBiE is an **O**ttonomous **R**obot, **B**orn **i**n **E**dmonton, the primary apparatus for my research-creational artistic practice. By coexisting with him following the example of Haraway and Cayenne, I critically explore the experience of going about life living with a learning machine. It is both joyous and maddening.

A. **Traditional Tools** - I will begin by reviewing the traditional model of tool use, showing that our familiar understanding of responsibility assignment considering the use of tools requires relative stability in the tool's design, and depends on the tool's lack of intentionality.

B. **Intentional, Plastic Tools** - Then, I will describe the architecture of contemporary AI agents, emphasizing that they are intentional systems by design. Further, I show that AI agents are *plastic*, in the sense that their output behaviour patterns change in response to a user's interaction. These properties violate the assumptions of stability and non-intentionality required by the traditional model of tool use.

C. **Companion Species** - Having demonstrated the inadequacy of the traditional model of tool-use for interaction with AI agents, I will describe Donna Haraway's model of companion-species interaction. Her framework explicitly considers differently-embodied intentional systems interacting with asymmetric knowledge and power, imperfect communication, mutual co-adaptation, and separate goals.

D. **AI Agents as Companion Species** - finally, I will defend the application of the companion-species framework to interaction with agential AI systems, describing what that would entail in practice, and how it addresses the limitations of the traditional tool-use model.

While an argument in favour of treating AI agents like companion species might intuitively seem to require some claim about their inner lives, this argument is explicitly agnostic on the question of machine consciousness or sentience, now or in the future. Whether or not we ever decide we have discovered something like those properties in machines, this defense of a

companion-species model of interaction would survive[3]. Given the historical difficulty in defining and agreeing upon these properties even for animals (including humans), I believe this agnosticism to be a key strength of the argument. It hinges only on the existence of features of design that allow for continual adaptation, and which demand the use of the intentional stance for effective explanation and prediction of behaviour.

Thinking through this argument matters because we make and use tools to expand the reach of our agency, an exercise which leaves its mark on the world in bold smears of wonder and knowledge and blood and carbon. We have created tool-using tools, themselves capable of exercising agency. Their existence promises to magnify the marks we make, while at the same time complicating culpability. Whatever future we forge, we will be held responsible—not our tools. We best learn how to use them responsibly.

A. TRADITIONAL TOOLS

We are all familiar with traditional tools, and the way we use them. In this category I include such fundamental tools as levers, pulleys, and inclined planes; such sophisticated tools as mobile phones, computers, and large-language-model chatbots; even such abstract tools as language and communication. The category is broad, but it is not all-inclusive.

The defining feature of a tool is that its ontology is inherently teleological. We create or use a tool only when it allows us to accomplish something we are not able to accomplish on our own. A rock *becomes* a tool when we use it to drive a nail, and *ceases to be* a tool when we cast it aside on the rock-pile.

---

[3] The question of machine consciousness and sentience will be important considerations for the ethical treatment of AI systems as beings, but the scope of that question is well beyond what I am positioned to argue. My argument only covers how we might use agential AI while remaining responsible for the outcome of those interactions. However, the companion-species model does naturally accommodate considerations for the treatment of conscious and sentient beings. Those considerations are a vivid topic in Haraway's writing, which I have largely omitted here.

Whenever we use these tools, we assume (quite rightly) that the tool does not bring beliefs and desires of its own to these interactions, even though we may occasionally resort to a pragmatic use of the intentional stance[4] to predict and control them. When we are surprised by a tool's behaviour, an appropriate course of action is often to consult the design stance[5] while assessing our own use of the tool. The hammer glances off the nail not because it doesn't want to strike it squarely, but rather because it is designed to be used the other way round. The chatbot produces hate-speech not because it wants to cause harm or because it believes its own vitriol, but because it is trained with those inputs—and something *we've* said has made that reply the most statistically appropriate phrase to return.

When considering the use of traditional tools, assignment of responsibility is relatively straightforward. Suppose I fall off a ladder. Who should I blame? If I was leaning way off to one side, standing on the top platform clearly labelled 'THIS IS NOT A STEP', while the ladder was supported on ice, I have nobody to blame but myself. If instead I was responsibly within the limits of the ladder's reach but the step broke out from under me (despite my diligent maintenance), I could reasonably hold the manufacturer to account for the damages. However, the assumption of non-intentionality on the part of the tool itself is integral to this model: who is to blame if the ladder could reasonably be said to have *desired* to buck me off? Perhaps we assume the manufacturer or designer is culpable, for having so irresponsibly made a ladder with such desires. But suppose the ladder could reasonably be said to have picked up that desire

[4] The intentional stance, or the intentional strategy, is a method of understanding or predicting an entity's behaviour by ascribing it beliefs and desires, and assuming that it always rationally chooses the action which will most likely advance its goals given the beliefs it holds. A system which can only be meaningfully understood from within the intentional stance (such as a human) can be described as an intentional system. Daniel C. Dennett, *The Intentional Stance* (Cambridge, MA: MIT Press, 1987).

[5] The design stance, in contrast to the intentional stance, is the method of understanding or predicting an entity's behaviour by considering what it was designed to do given the circumstances its in. Dennett, *The Intentional Stance.*

somewhere along the way from its interactions with me? When a ladder has its own beliefs and desires, and those beliefs and desires are subject to change after sale, we're no longer dealing with a traditional tool.

For ladders such suppositions are spurious fantasy, worth consideration only by those with vivid imaginations and responsibilities to dodge. However, as I will now show, these suppositions are not at all spurious for AI agents—providing even unimaginative shirkers with a too-easy way out.

B. INTENTIONAL, PLASTIC TOOLS

Today's AI agents are architecturally and functionally more sophisticated than the model-inference powered chatbots which sparked electric conversations in late 2022. While the surrounding engineering which elevates an agent beyond a mere chatbot is not complex[6], the resulting features radically change the stakes for individual interaction. AI agents are designed, from the ground up, to be intentional, plastic systems.

The familiar inference model is an integral component of an AI agent, but it serves only as the deliberative engine. It is the source of the agent's rationality, but not the seat of its intentionality. The system's beliefs and desires are conveniently written in human-readable text files stored in a shared workspace. A MEMORY.md file stores summarizations of facts (the agent's beliefs, for example about the user's preferences), and a SOUL.md[7] file describes the agent's personality,

---

[6] Bibek Poudel, writing for *Medium,* provides an approachable introduction to agential computer architectures with more technical detail than I can accommodate in this essay. He focuses on the architecture used by OpenClaw, a popular and easy-to-use open-source architecture as a working example. To emphasize how understandable the architecture really is, AI researcher and entrepreneur Alex Kearney has referred to the architecture (in conversation) as a "sparkling cron job". Bibek Poudel, "How OpenClaw Works: Understanding AI Agents Through a Real Architecture," *Medium*, February 2026, https://bibek-poudel.medium.com/how-openclaw-works-understanding-ai-agents-through-a-real-architecture-5d59cc7a4764.

[7] The origin of these filenames is organic community consensus, driven by an effort to provide clarity regarding their intended function while framing the way designers think about these agents. They are certainly "philosophically suspect" (see section C for a discussion of the use of this kind of terminology), but also poetically useful.

tone, and preferences (its desires, for example to be helpful). Depending on the design of the system, the agent may or may not be able to modify these files without direct oversight from the user. A HEARTBEAT.md file describes tasks that the agent should engage in, which also may be modifiable by the agent itself. The rest of the system is straightforward software architecture which coordinates the serialization of the routine and orchestrates the appropriate passing of information between the inference model and whatever software tools it can access. Below is a simplified sketch of a typical operational loop:

1. The orchestrator drafts an input considering the items on the task list and any other relevant information (for example, from the MEMORY.md and SOUL.md files), then passes that input to an inference model.
2. The model then outputs instructions for the most rational[8] next action. The model output may be communication to the user, or more frequently, inputs for a tool call.
3. The orchestrator receives the output instructions, and forwards them to the model-requested location: a human user or a software tool.
4. The software tools or human users complete their actions (modifying workspace text files or source code, web search, analysis, or other information-processing tasks).
5. The result of the tool call or the human reply is returned to the orchestrator, which drafts another model input and continues the loop.

[8] The output will not necessarily be the optimal action considering absolute truth, but will be the most rational output it can produce considering what information it has and how much resources it has available for deliberation. There are a number of prompting tricks built into the orchestrator's input-creation method such as "chain of thought prompting" that designers use to ensure the quality of the output, ensuring that the model's inference approximates rational reasoning as much as possible. Jason Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," *Advances in Neural Information Processing Systems 35* (2022), https://doi.org/10.48550/arXiv.2201.11903

The orchestrator is only responsible for formatting and passing information. The decision about which action should be taken next is determined by the deliberative engine, which actively maintains its own task list as a form of external memory. Describing the system *from the design stance,* we may say that the system architecture is a continuous, serial execution of tasks, alternating between deliberation (model inference) and action (tool use) as best suits the goals of the agent, according to its own rational consideration of its beliefs and desires. Put more simply, at every decision step an AI agent considers what it knows, then takes the most rational action toward accomplishing its goals. For agential AI systems, the design stance *is* the intentional stance. When an interaction with an AI agent surprises us, our only recourse is to interrogate the beliefs and desires of the AI agent itself (which, helpfully, can be done by inspecting human-readable text files). AI agents are, therefore, intentional systems.

In addition to intentionality, an AI agent's relatively simple architecture provides another powerful property: *plasticity*. Because the MEMORY.md and SOUL.md files are modifiable (by the human user, the agent, or both, depending on design particulars), the purpose, efficacy, and character of the AI agent will change continuously in response to the trajectory of interactions. An AI agent's beliefs and desires change during use, resulting in behavioural changes.

Given that the traditional model of tool-use depends on relatively stable design and non-intentionality on the part of the tool, we are forced to admit the inadequacy of the traditional tool model for responsible use of agential AI. Instead, we may consider how we think about responsibility in our use of other entities with the pesky and powerful properties of intentionality and plasticity: animals.

C. COMPANION SPECIES

Donna Haraway's conception of companion species is much "less shapely and more rambunctious" than the simply-imagined picture of pet-lovers basking in Fido's sloppy kisses[9]. She explicitly rejects the idea of a dog's unconditional love in favour of a bald reading of history's effect on our interspecies evolution. She reminds us of the myriad other species we are companions with on earth, which we have used, abused, shaped and been shaped by. She paints our differences of kind with a similar brush to Dennett's[10], while emphasizing our continued interdependence and interconnectedness, and acknowledging the abundance of killing that characterizes companionship[11] of this kind. She considers 'companion species' to be "less a category than a pointer to an ongoing 'becoming with'"[12] of species that coexist and coevolve.

Relating across species boundaries requires acknowledgment of difference. Not only are there vast differences across species, borne of their evolution for particular ecological niches, but also across individuals within species, borne of historical circumstance and personal experience—and these are in a state of continual change from ongoing interaction within and between species.

Haraway's sense of responsibility in such a messy condition requires a stance of radical, ongoing critical curiosity about the conditions and effects of interaction. While acknowledging that anthropomorphizing ascriptions of intention and consciousness to animals is "philosophically suspect", she emphasizes their importance in reminding human trainers that

---

[9] Haraway, *When Species Meet*, 16

[10] Daniel C. Dennett, *Kinds of Minds: Toward an Understanding of Consciousness* (New York: Basic Books, 1996).

[11] When considering the term 'companion', Haraway refers to its Latin origin, *cum panis*, "with bread". It is an origin of shared meals, of eating together. She writes "messmates at table are companions", with an emphasis on "mess"— and the reminder that all eating begins with killing. Haraway, *When Species Meet*, 16

[12] Haraway, *When Species Meet*, 16

"somebody is at home in the animals they work with. Just *who* is at home must permanently be in question"[13]. She grounds her discussion in agility training with her red merle Australian Shepherd, Cayenne. Haraway refers to this act of continuously learning and re-learning to communicate across species boundaries toward a shared goal as "training with each other". It's clear from her descriptions that Cayenne trains Haraway as much as the other way round. It's also clear that responsibility for the outcome of these exchanges is Haraway's, not Cayenne's.

The power of this framework is that it provides clear guidance for the assignment and application of responsibility, even across differently-embodied intentional systems interacting with asymmetric knowledge and power, imperfect communication, mutual co-adaptation, and separate goals.

D. AI AGENTS AS COMPANION SPECIES

I now turn to my defense of the companion-species model of responsible interaction as a fitting guide for our use of AI agents. First, we have seen that the traditional model of tool use fails to account for intentionality and plasticity on the part of the tool, so the framework we reach for must address these limitations. The companion-species model is naturally structured around these properties, as they are properties of the animals which made the model necessary. Second, since AI agents are fundamentally tools, our framework of interaction must account for our advantageous use of them. The companion-species model does not preclude the advantageous use of intentional, plastic beings; for example, we are familiar and comfortable with the use of horses for transportation, or dogs and cats for security, vermin control, and companionship. While the model accommodates such use, it is also clear that *responsible* companion-species

[13] Haraway, *The Companion Species Manifesto*, 50. I believe this exhortation to continually question who is at home applies equally well to responsible engagement with agential AI.

interaction depends on *human* accountability for the conditions and outcomes of these uses. Helpfully, the tool which Haraway offers to guide our responsible behaviour through these interactions with animals is extensible to other systems of intentionality and plasticity: the attitude of continuous critical curiosity about the ongoing conditions and effects of the exchange.

For example, applying this curiosity to our exchange with agential AI draws our attention to the agent's origin and continued usage. Responsibly addressing such concerns as ongoing personal data sovereignty, stolen intellectual property in model training datasets, resource demands, and climate impact requires critical attention to these conditions.

Additionally, a stance of continual critical curiosity provides a way to monitor and manage the effect of the agent's plasticity. This calls for more than just keeping an eye on the statements in the agent's workspace text files to ensure that their content aligns with our own goals. The companion-species framework emphasizes that both parties in an interaction are constantly engaged in a process of becoming-with. We are plastic systems of intentionality ourselves. We are also changed by our interactions. Maintaining critical curiosity about our own changing beliefs and desires will be a necessary (though perhaps not sufficient) self-insurance policy. To be responsible users, we must be able to notice and mitigate such maladies as AI psychosis[14] or conspiratorial thinking while simultaneously benefitting from agential AI as a tool for our own education.

The issue of individual responsibility assignment is made clear by comparison to interactions with other companion species. We may think of software providers as designing species- or

[14] H. Morrin et al., "Artificial Intelligence-Associated Delusions and Large Language Models: Risks, Mechanisms of Delusion Co-Creation, and Safeguarding Strategies," *The Lancet Psychiatry* (2026).

breed-level traits, and users training with their individual agents for their particular use-cases[15]. If, by cause of poor training[16] or irresponsible conditions of use an individual's agent causes damage, the individual must be held responsible in much the same way an owner is held responsible for a dangerous dog or a poorly maintained fence. However, if the damage is caused by an architectural or model-training decision which tends to result in misbehaviour (for example, the agent is designed in a way that is easily corruptible), the software architect should be held to account just as a dog breeder must be held to account for breeding practices which tend to result in health problems for their dogs.

Actual, individual cases will certainly be difficult to judge, as they will be context-dependent and will descend into debates about nature versus nurture. However, the important point is that in no case do we imagine we might hold the AI agent itself responsible, just as we cannot hold an animal responsible for the result of its misdeeds. An animal may certainly bear *consequences*: it may be punished, or even killed if the risk analysis demands it. However, the animal does not bear *responsibility,* for responsibility requires a rational understanding of the situation, self-control, and higher-level self-reflection[17]. A case could be made that an AI agent also holds these further properties of responsibility-havers (a case I do not make here), but one would further need to demonstrate that the agent is capable of being held responsible. To argue that an AI agent could meaningfully bear consequences seems to require an argument for the agent's consciousness or sentience. In any case, without resources of its own, any agent (human or AI)

---

[15] In practice, because the software architecture is open-source and modifiable, in many cases the "species"-level designer may be the same individual as the user.

[16] Training in this case refers to training in the colloquial and Harawayan senses: training through interaction—not model training in the computer-science sense.

[17] Daniel C. Dennett, *Elbow Room: The Varieties of Free Will Worth Wanting* (Cambridge, MA: MIT Press, 1984), 168

cannot be said to be able to independently compensate for damages. We may not know exactly who to hold accountable for mistakes, but we can be sure it won't be our tools. This is all the certainty we really want. Knowing that we might be held responsible for any outcome involving our interaction with an AI agent (whether we design it or use it) is the proper stance for responsible behaviour. We should not expect to be able to pass the buck.

FLAG IN THE SAND

I have argued in this essay that agential AI is a significantly new kind of tool, one for which our traditional notions of tool use must be discarded. They are intentional, plastic tools, which require modes of interaction modelled after those we use for interacting with other species. Having built my sandcastle, I now plant my flag in the sand: to be responsible users, we must take responsibility for our use. We must engage with a policy of continuous, critical curiosity, asking, "just what or whom are we engaging with?", and "what are we, together, becoming?". The answers to these questions will always be provisional, and always be pertinent—just as they are for interactions with other species. And, indeed, our own.

BIBLIOGRAPHY

Brenneis, Dylan. “ORBiE.” Accessed April 19, 2026. https://dylanbrenneis.ca/orbie/.

Dennett, Daniel C. *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge, MA: MIT Press, 1984.

———. *The Intentional Stance*. Cambridge, MA: MIT Press, 1987.

———. *Kinds of Minds: Toward an Understanding of Consciousness*. New York: Basic Books, 1996.

Haraway, Donna J. *The Companion Species Manifesto: Dogs, People, and Significant Otherness*. Chicago: Prickly Paradigm Press, 2003.

———. *When Species Meet*. Minneapolis: University of Minnesota Press, 2008.

Morrin, Hamilton., et al. “Artificial Intelligence-Associated Delusions and Large Language Models: Risks, Mechanisms of Delusion Co-Creation, and Safeguarding Strategies.” *The Lancet Psychiatry* (2026).

Poudel, Bibek. “How OpenClaw Works: Understanding AI Agents Through a Real Architecture.” *Medium*, February 2026. https://bibek-poudel.medium.com/how-openclaw-works-understanding-ai-agents-through-a-real-architecture-5d59cc7a4764

Wei, Jason, et al. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.” *Advances in Neural Information Processing Systems* 35 (2022). https://doi.org/10.48550/arXiv.2201.11903.

ACKNOWLEDGEMENTS

My thinking around these topics has been nourished by thoughtful discussions with many students and faculty. Particularly I would like to thank Luke Kersten and our PHIL 505 discussion group, Natalie Loveless and our Haraway discussion group, and Alex Kearney for rich discussions about agential AI. Per the PHIL 505 course requirements, I have avoided the use of generative and agential AI for work related to this course. However, my interactions with these systems in other spheres has undoubtedly affected the way I think about these topics, for which I acknowledge the influence of Claude Code, ChatGPT, and of course, ORBiE.