Sara M. Watson, *Data Doppelgangers and the Uncanny Valley of Personalization*

"What is it about my data that suggests I might be a good fit for an anorexia study?" That's the question my friend Jean asked me after she saw this targeted advertisement on her Facebook profile:

She came up with a pretty good hypothesis. Jean is an MIT computer scientist who works on privacy programming languages. Because of her advocacy work on graduate student mental health, her browsing history and status updates are full of links to resources that might suggest she's looking for help. Maybe Facebook inferred what Jean cares about, but not why.

Days later, I saw a similar ad. Unlike Jean, I didn't have a good explanation for why I might have been targeted for the ad, which led me to believe that it could be broadly aimed at all women between the ages of 18 and 45 in the greater Boston area. (When I clicked to learn more about the study, this was listed as the target demographic.)

Still, it left us both with the unsettling feeling that something in our data suggests anorexia. Ads *seem* trivial. But when they start to question whether I'm eating enough, a line has been crossed. I see similar ads plastered across the Boston T recruiting participants for medical studies on diabetes, bipolar disorder, and anxiety, but their effect is materially different. The only reason I see those ads is because I ride the T. These messages offer the opportunity to self-select for eligibility. It's different online, where I am supposed to see ads because something about my data suggests that they are relevant to me.

Personalization purports to be uniquely meaningful, yet it alienates us in its mass application.

Google thinks I'm interested in parenting, superhero movies, and shooter games. The data broker Acxiom thinks I like driving trucks. My data doppelgänger is made up of my browsing history, my status updates, my GPS locations, my responses to marketing mail, my credit card transactions, and my public records. Still, it constantly gets me wrong, often to hilarious effect. I take some comfort that the system doesn't know me too well, yet it is unnerving when something is misdirected at me. Why do I take it so personally when personalization gets it wrong?

Right now we don't have many tools for understanding the causal relationship between our data and how third parties use it. When we try to figure out why creepy ads follow us around the Internet, or why certain friends show up in our newsfeeds more than others, it's difficult to discern coarse algorithms from hyper-targeted machine learning that may be generating the information we see. We don't often get to ask our machines, "What makes you think that about me?"

Personalization appeals to a Western, egocentric belief in individualism. Yet it is based on the generalizing statistical distributions and normalized curves methods used to classify and categorize large populations. Personalization purports to be uniquely meaningful, yet it alienates us in its mass application. Data tracking and personalized advertising is often described as "creepy." Personalized ads and experiences are supposed to reflect individuals, so when these

systems miss their mark, they can interfere with a person's sense of self. It's hard to tell whether the algorithm doesn't know us at all, or if it actually knows us better than we know ourselves. And it's disconcerting to think that there might be a glimmer of truth in what otherwise seems unfamiliar. This goes beyond creepy, and even beyond the sense of being watched.

We've wandered into the uncanny valley.

* * *

Since the 1970s, theorists have used the term "uncanny valley" to describe the unsettling feeling some technology gives us. Japanese roboticist Masahiro Mori first suggested that we are willing to tolerate robots mimicking human behaviors and physical characteristics only up to a point: When a robot *looks* human but still clearly isn't.

The threshold is where we shift from judging a robot as a robot and instead hold it against human standards. Researchers at the University of Bolton in the UK have described this shift as the "Uncanny Wall" in the field of digital animation where increasing realism and technological advancements alter our expectations of how life-like technologies should be. I would argue that we hit that wall when we can't distinguish whether something is broadly or very personally targeted to us. The promise of Big Data has built up our expectations for precise messaging, yet much of advertising is nowhere near refined. So we don't know how to judge what we are seeing because we don't know what standard to hold it against.

The uncanny valley of robotics is grounded in the social cues of the visual. We are repulsed by the plastic skin, by the stilted movements, by the soulless eyes of our robotic counterparts. In contrast, personally targeted digital experiences present a likeness of our needs and wants, but the contours of our data are obscured by a black box of algorithms. Based on an unknown set of prior behaviors, these systems anticipate intentions we might not even know we have. Our data may not be animate or embodied like a robot, but it does act with agency. Data likeness can't be seen or touched, but neither can our sense of ourselves. This makes the uncanny even more unnerving.

Uncanny personalization occurs when the data is both too close and not quite close enough to what we know about ourselves. This is rooted in Sigmund Freud's famous treatment of the uncanny, which he traced to the feelings associated with encountering something strangely familiar. In Freud's original writing, the uncanny is the *unheimlich*—literally translated as "unhomely," and the opposite of *heimlich*, which is the familiar, comfortable feeling of being at home.

Technologies that are simultaneously familiar and alien evoke a sense of dread. In the same way, when our data doesn't match our understanding of ourselves, the uncanny emerges. Freud explains that *heimlich* also means that which is kept hidden, as in the private sense of the home. So when something becomes *unheimlich*, what should be hidden is exposed. We might think of our browsing history this way. With digital traces assembled by personalization engines, our most intimate behaviors are uncovered and reflected back to us. We don't think an ad is relevant to us, but it repulses us because we are worried that it could be.

A friend's Facebook status update captures this idea well: "I am never quite sure if Facebook's advertising algorithms know nothing about me, or more than I can admit to myself."

* * *

In his exploration of the uncanny, Freud also delves into the idea of the double. The doppelgänger is the identical other, and in the literature Freud cites, doubles are often connected almost supernaturally, sharing feelings, behaviors, and actions. Robots act as our mechanical doppelgängers, especially when they are designed in our likeness. Similarly, our digital data echoes our actual tastes and activities, often with higher fidelity than our own memories can. The familiarity of our data doppelgänger is uncanny—as though we are able to see our own body at a distance. When a person gets ads for "Flirty Plus-Sized Outfits," she might wonder, "does my browser behavior make me look fat?" The doppelgänger invites the strange possibility for self-observation and self-criticism. If we observe our double and we don't like what we see, we worry that the reflection might be more real than our perceptions of our actual selves.

Data isn't the first tool for self-reflection to produce a sense of the uncanny. The original reflective technology, the mirror, gave us the ability to see ourselves as others see us. French psychoanalyst and theorist Jacques Lacan described our fraught relationship with the mirror in the story of a man who sees his own back in the mirror and feels the presence of the ghost, unable to face himself in the reflection. The uncanny thus emerges in the act of seeing oneself as the other. (Strangely, one's "true" reflection—as seen in mirrors designed not to reverse what they reflect—can inspire the same reaction.) Personalization holds up a data mirror to the self, collapsing the distance between subject and object, and yet it's impossible for us to face our data doppelgänger with complete knowledge.

Kate Crawford, a researcher at Microsoft's Social Media Collective, wrote in *The New Inquiry* that the lived reality of Big Data is "suffused with a kind of surveillant anxiety—the fear that all the data we are shedding every day is too revealing of our intimate selves but may also misrepresent us..." and that the "anxiety of those surveilled is deeply connected to the anxiety of the surveillers." This is another sort of doubling. I worry about what Facebook knows, but Facebook also compulsively asks me for more information:

Facebook always wants to know more information about important people and memorable moments in my life. The "life events" that are most valuable to advertisers are often the most emotionally charged on a personal level. That's why the Target data team puts so much work into identifying the buying patterns of women likely to be in their second trimester of pregnancy. It's also why life-event advertising can go so horribly wrong when it is misfired, as it was in a Shutterfly marketing debacle a few weeks ago. Shutterfly sent a mass email to a broader-than-intended group of recipients, congratulating them on their new arrival. But not all of those who received the targeted email had newborn babies. While this slight was amusing to the friend who sent it to me, a 55-year-old woman with two grown boys, it was heart-wrenching to those sharing on Twitter that they were reminded of fertility problems or recent miscarriages.

Data approaches the uncanny valley, too, when it touches on death. Take the case of Mike Seay, who discovered that data brokers were keeping track of the death of his daughter when he

received a piece of mail with that information mistakenly printed between his name and his address. Printed on the envelope: "Mike Seay, Daughter Killed in Car Crash, Or Current Business."

Deaths in the database haunt us like the "*unheimliches* house" that Freud described. Interactions with advertisers used to feel simpler. When you bought a parenting magazine, you were opting to join a certain demographic that went along with that media. Today, simply googling a parenting question will lump you in with that demographic, regardless of whether you are or want to be identified as part of it. Personalized digital experiences are increasingly granular and niche. They hit closer to home, and they often confront us unexpectedly.

The only way to assuage our anxiety about the uncanniness of personal data is to develop causal explanations that link our digital experiences with the data they are based upon. Facebook just took a noteworthy step toward letting users know why they see certain targeted ads. The platform says it will allow users to click on an ad to find general explanations about the sources that Facebook uses to develop marketing profiles for advertisers.

But this is only a partial transparency. Even in the press release, Facebook offers only "one of the reasons" you see an ad, and this explanation oversimplifies the complexity of the data points involved in producing an ad. Such transparency is insidious because it uncovers some uses while further obscuring others. This gap is important because, for now, ads are our main insight into the collection and use of our personal data. They are the most legible instantiation of our data. And the same data that is used in an attempt to sell you merchandise could be used to influence the outcome of a loan decision or algorithmically determine citizenship.

As our behaviors, bodies, and environments are made legible as data, and as our online experiences mesh with our offline ones, we need to try to unpack these uncanny encounters with data. Throughout history, new technologies provoke moral panic and anxiety—in part because those technologies upend our understanding of time, place, and ourselves. But as we adopt and domesticate them, these technologies become integrated into our lives and embedded in the cultural fabric. The more time we spend time with our data doppelgängers, the more familiar they may become. That's why it is so important to be able to scrutinize our data and hold accountable the systems collecting our data while those processes are still malleable. The same dominant sociotechnical systems that favor data for its objectivity put our subjectivity at risk. We need to demand more ways to keep our data doppelgängers in check.