SAVANNAH
LAW REVIEW
VOLUME 5 | NUMBER 1

My Ex-Boyfriend Was an Automaton: Narcissism, Selfhood, and Discrimination in the Algorithmic Age

Margot Parmenter[*]

> "[B]eing *human* is synonymous with having feelings for other people . . . . The inverse of that . . . is the unfeeling double, the monster, doll, or robot: the alien other as mirror reflection."
>
> Siri Hustvedt, *The Delusions of Certainty*[1]

> "Or maybe this is a thing that he has, that he is, the same thing that more and more men have, some essential, awful quality of twenty-first century straight masculinity, and the end of romance: he's one of the evil fake."
>
> Kristin Dombek, *The Selfishness of Others: An Essay on the Fear of Narcissism*[2]

> "[T]hese distinctively minded beings are also a distinctive kind of self. Since they do lack consciousness, no (conscious) self would regard a zombie self as being equivalent to a normal self. Zombies claim to have feelings and sensations—along with conscious thoughts and memories—but all they really have are non-conscious states in their information-processing systems that they *call* 'feelings', 'thoughts', 'memories' and so on. A zombie will say that it is in pain if it breaks a leg, but in reality it feels nothing; the same goes for zombie

[*] Margot Parmenter is a graduate of the University of Oxford (Bachelor of Civil Law, 2016), Pepperdine University School of Law (J.D. *summa cum laude*, 2013) and the University of Chicago (A.B. in History, 2010). She clerked on the Court of Appeals for the District of Columbia Circuit and served as a Research Fellow at the Scuola Superiore Sant'Anna in Pisa, Italy. Her work focuses on the role that legal language plays in culture.

[1] SIRI HUSTVEDT, A WOMAN LOOKING AT MEN LOOKING AT WOMEN: ESSAYS ON ART, SEX, AND THE MIND 265 (2017).

[2] KRISTIN DOMBEK, THE SELFISHNESS OF OTHERS: AN ESSAY ON THE FEAR OF NARCISSISM 32 (2016).

> declarations of passion, or disgust, or outrage. Given all this, there are good reasons for regarding zombie selves—and lives—as possessing less intrinsic worth than normal conscious selves."
>
> Barry Dainton, *Self*[3]

Introduction

"My ex-boyfriend was an automaton." In 2018, artificial intelligence (AI) technology has not yet reached the stage of development where such a statement could be taken literally. Not, at least, without raising some important questions about definition and meaning. Instead, the statement is more naturally and immediately recognizable as a metaphor. When I deem my ex-boyfriend to be an automaton, I mean something that draws from the Siri Hustvedt and Kristin Dombek quotes above; something that mirrors the philosophical zombie problem articulated by Barry Dainton. In short, I mean to convey an idea that is familiar to us not only from the worlds of science fiction storytelling and horror movie myth, but also, increasingly, of modern cultural criticism—the notion that my ex-boyfriend is an unfeeling, emotionless shell; a simulacrum of a human capable of imitating humanity but not of having any; a being ultimately missing some important working "piece." Kind of like a narcissist—or, perhaps, a robot.

If we fast forward a few decades, however, it's possible that this statement may lose its metaphorical sheen. In a future of advanced artificial intelligence, we may find ourselves living alongside true automatons—Isaac Asimov[4] or *Ex Machina*[5]-style, really life-like, human-esque, Turing-Test-passing,[6] seemingly-

[3] BARRY DAINTON, SELF 194–95 (2014).

[4] Isaac Asimov wrote prolifically in the science fiction genre, and his "robot series" of short stories and novels—most of which are published in the collection entitled *I, Robot*—feature an imagined future where humanoid robots live alongside human beings. *See, e.g.*, ISAAC ASIMOV, I, ROBOT (1940–70), http://kaitnieks.com/files/asimov_isaac__i_robot.pdf. One of the most enduring legacies of these works is Asimov's Three Laws of Robotics, which constrain robot action by requiring: (1) that robots not injure humans or, through inaction, allow them to come to harm; (2) that robots obey orders given to them by humans, except where those orders conflict with the first law; and (3) that robots protect their own existence so long as such protection doesn't conflict with either the first or second laws. *See generally Do We Need Asmiov's Laws?*, MIT TECH. REV. (May 16, 2014), https://www.technologyreview.com/s/527336/do-we-need-asimovs-laws/.

[5] The 2015 movie, *Ex Machina*, featured an incredibly life-like humanoid robot who arguably achieves consciousness. *See, e.g.*, Angela Watercutter, Ex Machina *Has a Serious Fembot Problem*, WIRED (Apr. 9, 2015), https://www.wired.com/2015/04/ex-machina-turing-bechdel-test/; Martin Robbins, *Artificial Intelligence: Gods, Egos, and* Ex Machina, THE GUARDIAN (Jan. 26, 2016), https://www.theguardian.com/science/the-lay-scientist/2016/jan/26/artificial-intelligence-gods-egos-and-ex-machina.

[6] The Turing Test is father of computing Alan Turing's "famous speculation about discerning man from machine," in which "human judges are asked to hold an epistolary conversation with an entity . . . and attempt to discern whether that entity is a human or a computer program." ED FINN, WHAT ALGORITHMS WANT: IMAGINATION IN THE AGE OF COMPUTING 79 (2017). Though generally referred to in popular culture and scholarly discussion as a touchstone for the notion of detecting machine intelligence,

sentient[7] robots. For some people, this projected future might be a dream come true. For others, it might be a vision of the apocalypse. But for many of us jilted ex-girlfriends out there, such a future doesn't actually look that different from the present day. For us, those future automatons are just another variation on a type we already know all too well: with their seemingly-human features and their ability to appear (up until the critical moment) to be real and sentient—and with the difficulty they pose of discerning the real ones from the "evil fake[s]"[8]—these automatons have a lot of the same attributes and pose a lot of the same problems as our bad ex-boyfriends in the here and now. The automatons of the future and the ex-boyfriends of the past, then, present similar ontological and epistemological problems: in both figures, we confront the possible absence of some interior working piece, some essential inner spark that strikes us as intuitively necessary to the projects of communal, Earth-bound living: cohabitation, cooperation, and consciousness. We find ourselves dubious as to whether either of these beings have "selves"; whether they possess the ipseity that we often reference as an important shorthand in drawing and justifying our shared standards of conduct for life in society.

The similarity of these two entities is illuminating for the project of exploring how legal structures might adapt to the changing technological realities of AI development. Mining this metaphor allows us to probe the nature and implications of a potentially emergent "robot rights" regime in a manner that highlights often-overlooked sociocultural and discursive dimensions of law-making. By examining the ways in which bad ex-boyfriends might resemble future automatons, we can both (1) interrogate the intuitions that drive our collective efforts to craft agreed-upon standards of conduct from essentially-contested philosophical concepts;[9] and (2) trace the ways in which sociocultural trends

the test itself is not unproblematic. As Ed Finn aptly puts it: "The Turing Test was in many ways a demonstration of the absurdity of establishing a metric for intelligence; the best we can do is have a conversation and see how effective a machine is at emulating a human." *Id.* at 182.

[7] The term "sentience" in modern philosophy typically denotes the generic capacity to have sensory experiences—the ability of an entity to be "capable of sensing and responding" to its surrounding environment. *See Consciousness*, STAN. ENCYCLOPEDIA PHIL., https://plato.stanford.edu/entries/consciousness/ (last revised Jan. 14, 2014). However, the term is sometimes used interchangeably with "consciousness," which can encompass a variety of more complex capacities, like a certain sense of interiority—Thomas Nagel's thing that it is "like to be a bat," for instance—or even the concept of selfhood altogether. Thomas Nagel, *What Is it Like to Be a Bat?*, 83 THE PHIL. REV. 435 (1974); *see also* DAINTON, *supra* note 3, at 72–85. Herein, I use the term to denote the broader sense: beings capable of having sensory experiences and subjective interior states. This is why, with regard to the imagined automatons of the future, I have added the qualifier "seemingly."

[8] DOMBEK, *supra* note 2, at 32.

[9] The notion of an "essentially contested concept" was originally formulated by philosopher W.B. Gallie to describe concepts that express normative standards but about which there are varying interpretations. *See* W.B. Gallie, *Essentially Contested Concepts*, 56 PROC. OF THE ARISTOTELIAN SOC. 167 (1955–56). According to Samantha Besson, the "correct application" of such concepts involves creating "disagreement over [their]

impact the legal structures we imagine and build. In the gap between the here-and-now and the not-too-distant future, our metaphors matter. The ways in which we make sense of ourselves and of the world today inform the ways in which we will solve the problems of tomorrow. If one of those problems involves the "Rise of the Automatons," [10] we would do well to unpack the meanings of our "automaton" metaphors before either those meanings or the robots get away from us.

This Article proceeds in four parts. Part I describes and enumerates the automaton ex-boyfriend metaphor, exploring the ways in which the experience of encountering future, seemingly-sentient automatons might resemble that of encountering disappointing ex-boyfriends in that both experiences (1) prompt similar psychological phenomena and (2) raise similar philosophical quandaries. Part II examines the sociocultural grounding of our responses to the psychological processes and philosophical questions introduced in Part I by exploring one of the early-twenty-first-century's predominating cultural discourses about selfhood—that of the ever-present, narcissistic bogeyman who lacks a self but nevertheless lurks amongst us. This Part draws our attention to the cultural tools at our disposal in the prevailing narrative-discursive frameworks for managing the philosophical-psychological quandaries either endemic to the human condition or emergent in historical evolution. Part III deploys the insights uncovered in Parts I and II toward a thought experiment revolving around the applicability of anti-discrimination law in a future world where artificial intelligence is a reality. By asking whether anti-discrimination law might apply to a future where humans and automatons live together—where the joke that lends this Article its title can be taken literally—this Part aims to prompt consideration of the ways that cultural forces shape the laws we make and of the on-going process of negotiating between open-ended concepts like "humanity" or "equality" and actionable legal standards that shape conduct and decision-making. Part IV expands upon this broader point about the discursive translation process in the creation of legal structures by drawing upon the automaton/ex-boyfriend analogy to show how the introduction of a novel legal problem can expose the indeterminate philosophical undercurrents underlying many of the legal precepts we so often take for granted.

I. Bad Ex-Boyfriends, Automatons, and the Horror of Encountering Another Consciousness

Much of the potential for true horror often exists in the finest of details. Case in point: imagine that tomorrow you woke up in a world in which everything was exactly the same as it is now, save for the fact that your partner's eyes have changed from blue to brown. Everything else about him[11] is the same: his hair, his ears, his eyelashes, his nose, his clothes—even his old, worn-down running shoes

correct application." Samantha Besson, *Sovereignty in Conflict* (2004), http://eiop.or.at/eiop/pdf/2004-015.pdf (internal citations and quotations omitted).

[10] The *Rise of the Automatons* colloquium, at which this paper was first presented, was held at Savannah Law School on September 15, 2017.

[11] Or her, as the case may be.

in the corner of the room. *You* are the same; your life together is the same—and yet. Imagine the terror building as you began to rule out reasonable explanations: no, he isn't wearing color contacts as a ploy to tease or scare you; no, he hasn't been to the eye doctor recently; no, there's been no eye-altering surgery or aneurysm with side effect or change in prescription eye drop—what are you talking about? Of course his eyes have always been brown; look at all the photos—from your wedding, the honeymoon, that vacation you all took as a family last summer. Imagine your heartrate racing frantically as the objective evidence piles up: all the pictures on your phone also feature this brown-eyed imposter; his mother insists that his eyes—like those of many other babies—changed from blue to brown in the early months of his life, and of course she would know what color eyes her own son has and by the way is everything all right with you, dear? You are sounding a bit strange. Soon enough you would be either (a) convinced that this man was not your partner and that you were the completely-sane victim of an elaborate gas-lighting plot; or (b) uneasily dubious of your own experience, ready to consider and assimilate into your worldview the notion that you have just been wrong about this thing, that his eyes really have been brown all along. (And if you are a lawyer or a jurist, you might be particularly liable to the second outcome, trained as you have been to compulsively follow the evidence).

But why is this particular fault line—between things as we know them or are accustomed to them and things just-ever-so-slightly *off* of that known default—so ripe with potential for the unsettling? We know that it is—after all, stories in this trope are one of the mainstays of the horror genre;[12] this kind of story not only accompanies a general preoccupation in the human imagination with an array of almost-but-not-quite-human creatures like zombies,[13] animated dolls,[14] and—last but not least—robots,[15] it also expresses itself in the documented psychological

[12] *See, e.g.*, INVASION OF THE BODY SNATCHERS (Walter Wanger Productions 1956); THE STEPFORD WIVES (Palomar Pictures 1975); CORALINE (Laika 2009). Also consider, for a more recent entry in the field, Carmen Maria Machado's short story "The Husband Stitch." Carmen Maria Machado, *The Husband Stitch*, GRANTA (Oct. 28, 2014), https://granta.com/the-husband-stitch/. The relevant just-a-bit-off tale is actually a story-within-a-story, about a girl on holiday in Paris with her mother who returns to their hotel room to find her mother vanished and "the walls a different colour, the furnishings different than her memory" and a "hotel clerk [who claims] he has never seen her before."

[13] Zombies are a fantasy mainstay, and the list of recent incarnations in the popular imagination is a long one. Perhaps the most recent touchstone is AMC's long-running show *The Walking Dead* (American Movie Classics 2010). There is also "significant academic discourse" dealing with zombies, including a particular strain in philosophy that uses the figure of the zombie to explore the notion of consciousness. Shawn H.E. Harmon, *Zombies! Not Just the Undead, But the Near-Dead and the Never-Living: An Introduction to SCRIPTed's "Zombie" Analysis Section*, 7 SCRIPTED 336, 336 (2010). The philosophical zombie debate is discussed in *infra* Part II.

[14] As Siri Hustvedt eloquently puts it: "Put a knife in the hand of an innocent-looking doll that can walk, and you have a horror movie." HUSTVEDT, *supra* note 1, at 266. Perhaps the most famous walking doll of horror is Chucky, who first featured in *Child's Play* (United Artists 1988) and went on to appear in a variety of films.

[15] Although robots may be less frequent horror-movie characters, there have been some "frightening automatons" on film, like Ash from *Alien* (20th Centrury Fox 1979).

phenomenon of the uncanny valley, which describes the particular revulsion people experience when encountering a humanoid object.[16] Perhaps the idea of a beloved person or even a well-known thing being just-ever-so-slightly *off* strikes us as so horrific because we draw upon metaphor and narrative to make ourselves, and important questions about humanity, identity, and reality become more acute as the distance between categories contracts and the ease of differentiation recedes.[17] Thus, while it can be ostentatiously frightening to watch a Jason-masked slasher terrorize teenagers in a small town, it can be quietly but even more deeply horrifying to contemplate small differences between the world as we think we know it and the world as it might be—or between our own worldviews as shaped by the important characters who populate our worlds and the worldviews of those important characters themselves, as viewed from a perspective entirely external to our own.

In the contemplation of small differences, we encounter questions like: Who am I? Am I real? How can I be sure? What makes me, me? And: Are other people real? Are they like me? How can I be sure? Such questions, which force us to consider the very fundamentals of our lives and realities, harbor a certain existential terror—and it is this particular type of existential terror that typifies the encountering of both ex-boyfriends and seemingly-sentient automatons. Both such confrontations fall into the psychological-philosophical genre that I like to call "The Horror of Encountering a Separate Consciousness." Because of this affinity, they resonate similarly, provoking the same sorts of questions and problems and perhaps prompting us to reach for analogous narrative devices by which to assimilate experience and attribute meaning. How this is the case will become clearer after a more detailed exploration of each type of experience.

A. The Bad Ex-Boyfriend: He's a Human, But Does He Have a Self?

1. The Psychological

The story of the Bad Ex-Boyfriend is a story about desire and connection; it is about the ways in which we find—or maybe lose—ourselves in other people. There are, of course, as many variations on the theme as there are broken relationships, but, broadly speaking, the stories in this genre always have three acts: The Beginning, The End, and The Summation. The horror takes place mostly in Act III, when we are forced to confront the meaning that the dissolution

Erik Sofge, *The Automata of Terror: Cinema's 8 Scariest Robots*, Mental Floss (Oct. 30, 2017), http://mentalfloss.com/article/53453/automata-terror-cinemas-8-scariest-robots.

[16] The uncanny valley phenomenon describes the "characteristic dip in emotional response that happens when we encounter an entity that is almost, but not quite, human." It's a documented psychological response that can be triggered by "[a]nything with a highly human-like appearance." Stephanie Lay, *The Uncanny Valley: Why We Find Human-Like Robots and Dolls So Creepy*, The Conversation (Nov. 10, 2015), https://theconversation.com/uncanny-valley-why-we-find-human-like-robots-and-dolls-so-creepy-50268.

[17] *See generally* Derek J. Skillings, *Life is Not Easily Bounded*, Aeon (Oct. 24, 2017), https://aeon.co/essays/what-constitutes-an-individual-organism-in-biology.

of a partnership has for the reality of both the other's identity and our own (Now that everything is over, were they ever really *there*? And, am I still here? How can I be sure?). Nevertheless, it is important to take a glance at Acts I and II as well, because that is when the seeds for the horror that is to come are planted.

The Beginning of a new love is something that has been continuously mythologized by humans throughout time. What love is and how we experience it is a theme that has been explored by thinkers across disciplines for most of history. So, since it would be impossible to reduce that canon to a digestible synopsis, I will choose only a few representative accounts to describe this act, all of which express a common theme: the idea that new love involves, at turns, either the finding or the losing of oneself in another. Love is often conceptualized as a way to complete oneself, or perhaps as a way to transcend oneself, but either way the experience is couched mostly in first-person, subjective terms.[18] It seems that even in this most intimate and revelatory of all interactions, we are still bounded by our own ipseity.[19] We can see the explication of this notion in the work of three different thinkers across time and space—Plato, nineteenth-century French novelist Stendhal, and philosopher René Girard.

One of Plato's more enduring accounts is actually not a view on love that he endorsed; instead, it is the myth drawn by the character Aristophanes in the *Symposium*.[20] According to Aristophanes' account, humans were originally doubles of how they seem now—they had "four hands and four feet" and "one head with two faces."[21] Because this "primeval man was round, his back and sides forming a circle," he "could walk upright as men now do, backwards or forwards as he pleased, and he could also roll over and over at a great pace, turning on his four hands and four feet, eight in all, like tumblers going over and over with their legs in the air."[22] Though this sounds patently ridiculous (and likely did to those at the Symposium as well, as Aristophanes was ancient Athens's most famous comic playwright)[23] the mythical original humans were apparently so "terrible" in "might and strength" that the gods figured something needed to be done to control them.[24] So Zeus came up with a plan to cut the humans in half, and, as a result, each of us is now "but the indenture of a man, and he is always looking for

[18] Such a reality may, indeed, be embedded in the nature of ipseity, which Siri Hustvedt aptly describes as the "'for me' quality of conscious life." HUSTVEDT, *supra* note 1, at 241. Nagel's *What Is It Like to Be a Bat?* argument mirrors this notion, suggesting, as it does, that "the subjective experience of being you, me, or a bat takes place from a particular first-person perspective *for* you, me or the bat and that no objective third-person description can fully characterize that reality." HUSTVEDT, *supra* note 1, at 241.

[19] *Id.*

[20] PLATO, THE SYMPOSIUM (Benjamin Jowett, trans. 2009) (c. 360 B.C.E.), http://classics.mit.edu/Plato/symposium.html.

[21] *Id.*

[22] *Id.*

[23] Maurice Platnauer & Oliver Taplin, *Aristophanes*, ENCYCLOPEDIA BRITANNICA, https://www.britannica.com/biography/Aristophanes (last visited Apr. 19, 2018).

[24] PLATO, *supra* note 20.

his other half."[25] It is because of this, then, that people experience love as an "intense yearning" to "grow together, so that being two . . . become one."[26] According to Aristophanes, "[l]ove is born into every human being; it calls back the halves of our original nature together; it tries to make one out of two and heal the wound of human nature."[27] The magic of new love, by this account, derives from the operative force that breeds the desire to love in the first place—the desire to find oneself by locating the "completing" puzzle piece. By finding our missing half, we both transcend and perfect our deficient half-selves. Such an idea is not far from the more straight-forward description offered by nineteenth-century Scottish poet Alexander Smith, who wrote that "love is but the discovery of ourselves in others, and the delight in the recognition."[28] From Plato, then, we get the idea that love involves a certain transcendence—the dissolution of self into other, or the making of oneself through another. Either way, love involves a sort of subjective entangling of self with other.

Nineteenth-century French novelist Stendhal picks up this thread with the idea that the birth of love entails a psychological remaking of the beloved into a worthy and desirable being. Stendhal, in his essay "On Love," called the process "crystallization," and used it to describe the ways that we "endow" those we love "with a thousand perfections," in the end "overrat[ing] wildly by "draw[ing] from everything that happens new proofs of the perfection of the loved one."[29] As Noel Perrin later articulated it, Stendhal recognizes that "love is largely self-generated," the beloved being "less a person one meets than a person one creates."[30] In other words, the Beginning of love has little to do with any objective occurrence, and more to do with an internal transformation wrought in and through ourselves and only *projected* outward onto a love object. Perhaps this view of love best accompanies a notion of self-expansion—the idea that love gives us the ability to escape "the utter uncertainty of our situation"[31] by swelling beyond the anxiety that so often defines and confines our human condition. The notion of crystallization thus encompasses that familiar trope of new love's intoxicating happiness—the way it makes colors seem brighter, food taste better, and life

[25] *Id.*

[26] *Id.*

[27] PLATO, *supra* note 20; *see* Firmin DeBrabander, *What Plato Can Teach You About Finding a Soulmate*, THE CONVERSATION (Feb. 13, 2017), https://theconversation.com/what-plato-can-teach-you-about-finding-a-soulmate-72715.

[28] *Love Is But the Discovery of Ourselves In Others, and the Delight In the Recognition*, PHILOSIBLOG (Dec. 24, 2012), http://philosiblog.com/2012/12/24/love-is-but-the-discovery-of-ourselves-in-others-and-the-delight-in-the-recognition/.

[29] Faena Aleph, *How Does Love Crystallize? Stendhal Responds* (Aug. 19, 2012), http://www.faena.com/aleph/articles/how-does-love-crystallize-stendhal-responds/; *see also* STENDHAL, LOVE (Gilbert & Suzanne Sale trans., Penguin Books 1975).

[30] Noel Perrin, *The Heart and Its Reasons: Falling in Love with Stendhal* (Oct. 18, 1981), https://www.washingtonpost.com/archive/entertainment/books/1981/10/18/the-heart-and-its-reasons-falling-in-love-with-stendhal/5b9245ec-cf7b-414a-a18d-893286ce97eb/?utm_term=.86fed8414160.

[31] DeBrabander, *supra* note 27.

appear altogether more beautiful. Love, at the Beginning at least, elevates everything, including both the other and ourselves.

Finally, René Girard depicted the sort of self-delusive/self-escapist love project hinted at by Stendhal as "the ordinary dynamic of all desire."[32] According to Girard, "[w]e're all performing self-sufficiency as best we can" even though "we've become selves [in the first place] by imitating others" and "dependence on others is our fundamental, existential state."[33] But because we desire self-sufficiency, or perhaps because we feel elementally uncomfortable with the dissonance between our experiential ipseity and the fact that we only learn and define our selfhood through the involvement of others, we tend to "fall in love with people upon whom we can project our fantasies that there are some selves that are, unlike our own, replete unto themselves."[34] Of course, this projection proves illusory in the end, and when the bubble is burst, we are left to grapple with the incongruity between "the fullness we fantasized" and the emptiness it leaves in its wake.[35] When love ends, it is as if we are thrust back into ourselves—if we had found ourselves, we are now lost again; if we'd been able to blissfully lose ourselves, we are now stuck with them again.

In Act I (The Beginning) of the Bad Boyfriend Story, then, we can see the seeds of the later danger: love crosses—or, is the most serious, personally-meaningful attempt to cross—the ontological chasm erected by individuation. While it persists, we can ignore the problems of our own personhood; we need not contend with ourselves. We can find ourselves in the partner—the mirror of the other—and we need not assure ourselves of our own reality. Enter Act II.

Act II, the End, can happen bit by bit or all at once. But with the sort of bad boyfriend that we are particularly apt to diagnose with narcissism in the aftermath, it always involves what Kristin Dombek articulates as a kind of "turning away."[36] It is this turning away that makes Act II of the Bad Boyfriend story like a horror film: whether the end comes slowly or in one fell swoop, it's characterized by a change so minute that you can't tell whether it's real or whether you have suddenly gone mad. On his surface, everything is exactly the same as it was yesterday, yet he says that the most important thing has changed. The thing that was invisible to begin with, but of which you had previously been so sure, is gone: he doesn't love you anymore. As Dombek hauntingly puts it, "[t]he eyes that gazed upon you with such life, lit up by you, are now the dark stone eyes of a fake, made thing or an animal, turning away from you. [37] In response to this annihilation—this degradation that dismantles the world as you knew it while so cruelly leaving all the physical attributes exactly the same—you may start to feel indignant, and you will certainly be scared. Without his mirror there to reflect yourself back to you, you will wonder who you are. You will wonder whether you are real, and how to find yourself, or know yourself, or assure yourself of your

[32] DOMBEK, *supra* note 2, at 40.

[33] *Id.*

[34] *Id.*

[35] *Id.*

[36] *Id.* at 31.

[37] *Id.*

existence, without the completion or the reflection offered by the other. In this way, then, the end of love with a disappointing ex-boyfriend involves an intimate encounter—on very personal and impactful terms—with the reality of the other. In the dissolution that accompanies the end of a relationship, the thing that was once a unit dissolves. His turning away is a return to himself, a reassertion of his separate, distinct selfhood. But if he's a self again, then what are you? It is the attempt to answer this question in Act III that may find you transmuting your doubts about yourself into a denial of his—transposing, perhaps, your own feelings of loss and fear into a diagnosis that reveals him to be a showman, a simulacrum, one of the "empty fake."[38]

Act III, The Summation, is where you grapple with the psychological horror of having to confront this distinct other. It's where you decide what story you are going to tell yourself about him; how you are going to assimilate into your worldview (and self-view), this emergent, autonomous reality in which he acts by, of, and for himself—appearing both very familiar and terribly foreign. It's where you decide what you will do with your sudden, unexpected awareness that the other is fundamentally unknowable, incalculable, and unpredictable—even under the most intimate of circumstances. And it is where you decide what that awareness means for and about you, whether it reaffirms or calls into question your own self. In another kind of break-up, this debrief might diagnose the disappointing ex-boyfriend as simply "wrong for you," or "kind of a jerk," but with the bad ex-boyfriends we most often want to paint as hopeless narcissists, the resolution feels altogether more urgent and more critical. It harkens back to that horror scenario from the opening of Part I, and seems to split the world in two—it's a choice between the world as you know it, as you remember it (a reality in which your self is affirmed), and the world as it has been unexpectedly remade by someone else. Either he's real and you were wrong, or you're real and it was all a grand conspiracy. Indeed, because the confrontation of the bad ex-boyfriend's extrinsic selfhood engages our self-understanding and self-making on the most intimate of terms, The Summation in this case appears as the choice between him and you: either you downgrade yourself, abandoning your own memories of how his eyes were blue just the other day, or you unmask him as unreal—as lacking, as fake, as, for ease of reference, a narcissist. René Girard explains:

> We are quite ready to accuse others of 'narcissism' . . . particularly those whom we desire, with the aim of reassuring ourselves and relating their indifference, not to the very minor interest that we hold in their eyes or even perhaps in absolute terms . . . but to a kind of weakness that afflicts others. When we do this, we credit them with an excessive and pathological concentration on themselves—with a kind of illness that

[38] *Id.* at 55. This formulation derives from the thought of Girard, whose critique of Freud's "On Narcissism" suggested, according to Dombek's reading, "that our inability to get out of our own shoes when we encounter the selfishness of others can mean that what we end up diagnosing in them is our own fear and desire." *Id.*

> makes them more sick than we are and consequently incapable of . . . meeting us half-way as they should.[39]

At some point, when the break-up is farther behind you, you may be able to let his selfishness be only that—a return to him*self* and a turning away from you. In the meantime, though, you may be inclined to reach for a different narrative, a cultural "prescription for a pain reliever"[40] that explains how the Bad Ex-Boyfriend is actually an evil, "false self" who "masquerad[es] as real" only to "feed on other selves" in a sort of "contag[ion of] emptiness."[41] In this way, your psychologically horrifying encounter will lead you to a confrontation with one or some of the persistent questions of the human condition—philosophical questions about the self and the nature of being.[42]

2. The Philosophical

The story of the Bad Ex-Boyfriend involves a particular psychological phenomenon wherein we are forced to confront the reality of another, separate consciousness. It also raises certain accompanying philosophical questions. I've chosen to formulate these questions as ones about the nature of the self, and to conceptualize this reckoning process as one that questions what the self is and what it means to have one. In doing so, I've mostly followed the lead of essayist Kristin Dombek, whose insightful 2016 essay *The Selfishness of Others: An Essay on the Fear of Narcissism*[43] aptly captures the feelings and thoughts that accompany The Summation—the horror of being confronted with the reality of a separate consciousness in the form of a Bad Ex-Boyfriend.

When coping with this scenario, Dombek writes, we may begin to doubt whether our counterpart is altogether fully human—after all, how could someone who has all the working parts central to operating as a human in the world do what he did to me? It is at this point that—perhaps aided by culture, and by that dangerous rabbit-hole of information that calls itself the internet—it may occur to me that this boyfriend is indeed *not* fully human, that he is missing something which I and other "good" people have. Although he looks alright on the outside, perhaps there really is no *there* there; maybe he is empty inside. In selfhood terms, the doubt could be formulated something like this:

> Normal, healthy people are full of self, a kind of substance like a soul or personhood that, if you have it, emanates warmly from inside of you

[39] *Id.* at 39–40.

[40] *Id.* at 117.

[41] *Id.* at 11. The sociocultural problem that Dombek diagnoses in modern society, which is explored in more detail in Part II, is the prevalent tendency to pick a particular solution in this situation—to adopt a particular explanatory narrative by which we understand ourselves and others; namely, the narcissism script. *See infra* Part II.

[42] As Dombek recognizes in her explication of the cultural epidemic forming the central theme of her essay, which in this case manifests in a disappointing ex-boyfriend, our encounter with "the selfishness of others . . . leads quickly to the very difficult question of how we know things about others at all, and the mind-knotting question of how we know things at all." DOMBEK, *supra* note 2, at 12–13.

[43] *See generally id.*

> toward the outside of you. No one knows what it is, but everyone agrees that narcissists do not have it. Disturbingly, however, they are often better than anyone else at seeming to have it. Because what they have inside is empty space, they have had to make a study of the selves of others in order to invent something that looks and sounds like one. . . . It might take you a while to realize that the narcissist is not merely selfish, but actually doesn't have a self. When you do, it will seem spooky.[44]

But without a self, the bad ex-boyfriend quickly becomes something less than human; according to Dombek's formulation, "the narcissist is . . . a caricature of what we mean by 'not a good person' . . . [he's] a living, breathing lesson in what badness is."[45] So bad is this self-less,[46] fake simulacrum, that, indeed, the only thing to be done is to avoid him and all of his kind. And, if I confront the typical advice sources—the self-help blogs on the internet, my friends, the media, or pop culture dating books—I'll find that what I should do upon confronting one of these individuals, who are situated "outside the empire of normal mental health, flickering eerily at the edge of pathology," is simply to get away as fast as possible.[47] I should not engage, I should not try to understand, I should not feel sympathy; once I have seen the signs, once I have surmised that I am dealing with one of the "evil fake," I should simply "grab [my] running shoes and start [my] first 5K."[48] The self-less simulacrum should, in a sense, be cast outside of our shared community. He can't engage with us because he isn't enough like us.

Although this formulation does make sense of some of my psychological horror—relieving the dissonance between the conviction that I am real and my ex-boyfriend's newfound inability to pay me any regard—it doesn't all add up. Indeed, I might begin to wonder: "if he is empty inside . . . who or what is it, inside of him, that is imitating having a self? . . . Is he animating his selfiness with another, also fake part, of his selfiness? But what, then, is animating that part?"[49] Such questions problematize the narrative and disrupt the employed categories—what do we mean by a self, anyway? As applied to a Bad Ex-Boyfriend, can this concept really bear any meaning independent of other, value-laden judgements about how a person *should* be? About what's good and bad, acceptable or not? About how people should treat—and should be allowed to treat—one another? In short, could it be that the concept of the "self,"as deployed in this situation, is actually acting as a placeholder for a host of other issues? I would suggest that the answer is "yes," and would also direct our attention to the fact that these sorts of issues are the kind of thing the law typically settles (at least provisionally) for us. Thus, while the philosophical quandary of selfhood might seem very far away from the

[44] *Id.* at 6.

[45] *Id.* at 7–8.

[46] The term "self-less" is meant to denote those empty beings lacking a self, not to invoke the more typical meaning of the word that describes an attitude of charity.

[47] Dombek, *supra* note 2, at 9.

[48] *Id.*

[49] *Id.*

workaday aspects of the law, our discourses surrounding it are not altogether that distant from our legal structures.

The philosophical quandary raised by confrontation of the Bad Ex-Boyfriend, then, reveals something about the shape of our meaning-making processes. We navigate our psychological experiences using available extant categories and concepts, and we make and remake these concepts as we negotiate our experiences. In the case of the Bad Ex-Boyfriend, we (1) confront the reality of the other, and then (2) we try to work out what it means to live in that reality—to live in a world populated by other autonomous beings, that is not entirely of our own making—by negotiating the concept of selfhood. Understood in this way, it's possible to read the investigation of the self that accompanies confrontation of the Bad Ex-Boyfriend as an exploration of that concept which ultimately measures it in terms of another, perhaps more intuitive one—and that alternate concept is *humanity*. Because my ex-boyfriend is human, but my confrontation with him has prompted me to ask whether he's operating the way a person *should*, the selfhood discourse appears as a way to downgrade his humanity without denying it. This deployment of the concept is not, after all, unfamiliar—echoes of it appear in conversations surrounding psychopathy or sociopathy.[50] When people fall short of shared standards of conduct, we have a discourse for downgrading their humanness: they're human, but not entirely; something important is missing.

When it comes to the Bad Ex-Boyfriend, then, the confrontation of the other prompts a philosophical concern about the self. The ensuing discourse uses the concept of the self as a way of articulating and negotiating shared standards for being in the world. With the automatons of the future, the script is altered, but the underlying task remains the same: the confrontation of the seemingly-sentient automaton involves a formulation of the question of the self in terms of consciousness, where the conceptualization of consciousness—rather than the notion of humanity—stands in for our negotiation about the qualities required for being an entity in the world.

B. The Future Automaton: He's Definitely Not Human, But Does He Have a Self?

Encountering an Automaton that is virtually indistinguishable from a human being (except that we somehow know or suspect he may be a robot),[51] is similar to encountering a disappointing ex-boyfriend save for this: our opening intuitions about the case are reversed. I can't be sure that either being has a self, but with my Bad Ex I expected him to have one—I expected there to be a *there* there, and acted provisionally as if that were the case. It's only when he turned away that I started to become suspicious; up until that point I was happy to give him the benefit of the doubt. But when I imagine encountering the automatons of the future, this dynamic is turned on its head: assuming, as I do, that I will know that this

[50] HUSTVEDT, *supra* note 1, at 265–66.

[51] This is an important assumption that's embedded in the imaginative exercise itself, for the simple reason that without it the confrontation with the Automaton couldn't exist, rendering the exercise thereby insensible.

automaton is almost-but-not-quite human—that his artificial provenance will be marked out ever-so-subtly—I am not inclined to give him any such benefit of the doubt. Instead, skepticism is my default from the start. I go in expecting *not* to encounter a self, and am shocked when I feel as though I have encountered something resembling one. In the future-automaton scenario, I encounter selfhood seeming to emerge rather than seeming to withdraw, but the experience of horror is the same. Moreover, the accompanying philosophical discourse employs the same categories; but, because humanity is, by definition, off the table, it utilizes those categories in a different way. In response to this scenario, the concept of selfhood is used to explore the relevance of consciousness to capacity for participation in shared life.

To see how this is so, it will help to explore in more detail the conditions under which we might confront future robots.

1. The Psychological

In the projected future scenario wherein we confront seemingly-sentient automatons, we are liable to experience a similar sort of psychological phenomena—the existential horror of encountering another, separate consciousness. However, the contours of this experience vary in several respects from the case of the Bad Ex-Boyfriend. Broadly speaking, the differences track the fact that the encounter with automaton selfhood would be (1) emergent, in the sense that an entity which once acted only at human behest has begun to act of and for itself, and (2) alien, in that the selfhood of a sentient machine might be different in kind from our own. To give more shape to the possibilities of this projected future encounter, let's consider three examples: Ed Finn's reading of the modern algorithmic world, Isaac Asimov's fictionalized account of possible-robot Stephen Byerley, and the case of Cynthia Breazeal's "sociable robot" Kismet.

While we do not know exactly what it might be like to encounter a truly sentient robot, we need not look very far beyond modern life to gain insight into what it's like to encounter the workings of an intelligence different from our own. Indeed, as Ed Finn aptly articulates it in his work *What Algorithms Want: Imagination in the Age of Computing*, we already live alongside a sort of "algorithmic imagination" that we perceive as intelligent but lack the capacity to fully comprehend.[52] This intelligence is perhaps best exemplified by the machine learning algorithms that we increasingly depend upon to "parse complex data and make decisions."[53] While, "[g]iven time, data, and a precise statement of the problem, a machine learning algorithm can create a robust solution to . . . [a] problem," the process by which it combs and processes data strikes humans—even those immersed in the field—as "effective" and yet "inscrutable."[54] Thus, although machine learning algorithms have begun to penetrate daily life from the

[52] FINN, *supra* note 6, at 187–88.

[53] *Id.* at 182.

[54] *Id.* at 183.

most benign (recommending the next pilot on Netflix)[55] to the most sublime (it is around this field that technologists and other thinkers have begun to unite in pursuit of a "truly thinking machine")[56] we encounter the operation of those algorithms as a "kind of imagination at work," whose "flashes of complex order and process emerging from chaos" strike us as real but evade our comprehension.[57] Indeed, algorithmic decision-making pervades our reality, and "algorithmic machines now manage not just our individual streams of data but the great rivers of collective information, filtering our friends and colleagues as regularly as they curate our music or the news."[58] And the way we feel about these algorithmic machines is often marked by the same sort of revulsion-combined-with-attraction dynamic that marks all fascination with the uncanny.[59] Think of the feeling you have when targeted by an algorithmically-generated online advertisement that mirrors a Google search you made yesterday—it's the feeling of encountering something familiar and yet foreign: you remember running the search, but you no longer want tickets to that concert and it's sort of weird to be reminded of how much you once did. As Ed Finn articulates it, our modern encounters with algorithmic systems already possess some of the psychological horror of confronting an autonomous consciousness:

> [A]lgorithms have always embodied fragments of ourselves: our memories, our structures of knowledge and belief . . . . They are mirrors for human intention and progress, reflecting back the explicit and tacit knowledge that we embed in them. At the same time they provide the essential ingredient of mystery, operating according to the logics of the database, complexity, and algorithmic iteration, calculating choices in ways that are fundamentally alien to human understanding.[60]

According to Finn, this dynamic mimics a "love affair;" like with "every romance, the attraction [of algorithmic processing] is based on . . . the recognition of ourselves in another as well as the mystery of that other."[61] When I shop an algorithmically-recommended product, or watch a movie because of a Netflix prompt, I can see a dim reflection of myself—I remember watching other movies "with strong female leads"—but I also feel acutely aware that the algorithm is ultimately incalculable. In this way, the experience of the algorithmic world raises a psychological phenomenon similar in kind to that prompted by the throes of a bad break-up, with alien-ness substituting for dissolution in the generation of existential terror. AI, even at its still relatively modest modern stage of

[55] *See id.* at 183, 87–112.

[56] *Id.* at 182.

[57] *Id.* at 183.

[58] *Id.* at 186.

[59] DOMBEK, *supra* note 2, at 43–44.

[60] FINN, *supra* note 6, at 189–90.

[61] *Id.* at 189.

development, can leave us with the feeling that we've encountered a *there* there—an autonomous intelligence that can act of its own accord.[62]

Despite the magical, "ghost in the machine"[63] vibe of machine learning algorithms, our experience with these entities remains abstract in the sense that we still encounter them *as* machines or machine outputs. We don't yet encounter decision-making algorithms in embodied form, and it is not difficult to imagine that doing so would be a fundamentally unique experience, different in kind from interacting with, for instance, a chess-playing computer.[64] Transferring the projected encounter to a world populated by humanoid robots simultaneously heightens the drama and highlights the ways in which a confrontation with a future automaton might more closely resemble one with a Bad Ex-Boyfriend.

In Asmiov's classic text *I, Robot*, he tells the tale of Stephen Byerley, a consummately professional district attorney who may or may not be a robot.[65] In the story, Byerley runs for Mayor, and his opponent, a politician named Francis Quinn, becomes convinced that he is a "robot of a humanoid character," manufactured and maneuvered into place by the U.S. Robot & Mechanical Men Corporation as part of a plot to encourage the various Regions to allow the "use of humanoid . . . robots on inhabited worlds"—a state of affairs currently rendered impossible by public prejudice.[66] As Quinn puts it, "Suppose you get [the public] used to such robots first—see, we have a skillful lawyer, a good mayor, and he is a robot. Won't you buy our robot butlers?"[67] Convinced that he is right, Quinn spends the balance of the tale attempting to prove that Byerley is a robot—a feat in which he is ultimately frustrated because, as famous robot psychologist Dr.

[62] According to the Oxford Dictionaries, the word "automaton" derives from the Greek "automatos," which means "acting of itself." *Automaton*, OXFORD LIVING DICTIONARIES (2018), https://en.oxforddictionaries.com/definition/automaton.

[63] The term "ghost in the machine" was used by philosopher Gilbert Ryle to criticize Cartesian dualism, which, broadly speaking, holds that "each human being consists of a mind (which is a non-physical, purely spiritual thing) inhabiting a body, which is completely material and subject to the law of physics." Alan Brody, *Driving the Ghost from the Machine*, PHIL. NOW (1995), https://philosophynow.org/issues/13/Driving_the_Ghost_from_the_Machine.

[64] In the spring of 2017, the computer program *AlphaGo* famously beat the top world player at the ancient game, which involves "two contestants moving black and white stones across a square grid, aiming to seize the most territory." Rueter's News Agency, *Computer Beats Chinese Master In Ancient Board Game of Go*, THE TELEGRAPH (May 24, 2017, 1:23 AM), http://www.telegraph.co.uk/news/2017/05/24/computer-beats-chinese-master-ancient-board-game-go/. For a deeper explanation of the game of Go and of the program that learned to play it, see Christof Koch, *How the Computer Beat the Go Master*, SCIENTIFIC AM. (Mar. 19, 2016), https://www.scientificamerican.com/article/how-the-computer-beat-the-go-master/. *See also* David Silver et al., *Mastering the Game of Go with Deep Neural Networks and Tree Search*, 529 NATURE 484 (2016).

[65] ISAAC ASIMOV, *supra* note 4, *Evidence* at 144 (1946), http://kaitnieks.com/files/asimov_isaac__i_robot.pdf .

[66] *Id.* at 116–17.

[67] *Id.* at 117.

Susan Calvin explains it, a robot, constrained by Asimov's famous Three Laws,[68] would be indistinguishable from a good man.[69] Thus, while it is possible to prove that Byerley is *not* a robot, it's impossible to prove that he *is* one, a conundrum exasperated by the fact that the technology in Asimov's world is so advanced as to allow for a robot's physicality to appear exactly like a human's.[70] In this imagined iteration of the encounter, then, we have a problem of selfhood almost exactly identical to that at issue with the Bad Ex-Boyfriend, except for the fact that the search for the self in this instance is more explicitly a search for humanity driven by what Dr. Susan Calvin calls "a prejudice against robots which is quite unreasoning."[71] In other words, the need to prove that Byerley is a robot springs from nothing more than the meaning that categorizing him as a robot would have for him in the world. Encountering Byerley is, in all respects, exactly like encountering another human, but—because robots are subject to a certain prejudice and a different set of laws in Asimov's fictional universe—proving his robot provenance is important for what it means about the ways he will be allowed to operate in the world.[72] We thus have a glimpse, within this story, of a way in which the confrontation with a future, seemingly-sentient robot, although also prompting a psychological phenomenon grounded in the countenancing of a

[68] The reason for this is that "if you stop to think of it, the three Rules of Robotics are the essential guiding principles of a good many of the world's ethical systems." *Id.* at 121. A decent human would follow Rule Three because "every human being is supposed to have the instinct of self-preservation." *Id.* And any "human being . . . with a social conscience and a sense of responsibility" would "defer to proper authority. . . even when [it interferes] with his comfort or his safety," so a good human would also follow Rule Two. *Id.* Finally, a good human "is supposed to love others as himself, protect his fellow man, risk his life to save another," and so would follow Rule One. *Id.* As Susan Calvin concludes, then, "[t]o put it simply—if Byerley follows all the Rules of Robotics, he may be a robot, and may simply be a very good man." *Id.*

[69] Consider this scene:

> Quinn sat back in his chair. His voice quivered with impatience. "Dr. Lanning, it's perfectly possible to create a humanoid robot that would perfectly duplicate a human in appearance, isn't it?" Lanning harrumphed and considered, "It's been done experimentally by U. S. Robots," he said reluctantly, "without the addition of a positronic brain, of course. By using human ova and hormone control, one can grow human flesh and skin over a skeleton of porous silicone plastics that would defy external examination. The eyes, the hair, the skin would be really human, not humanoid. And if you put a positronic brain, and such other gadgets as you might desire inside, you have a humanoid robot."

Id. at 122–23.

[70] *Id.* at 123.

[71] *Id.* at 130.

[72] Not the least of these restrictions on operating within the world is the meaning that this categorization will have for whether Byerley is deemed—both by law and by public opinion—fit for holding elected office within the democratic polis. *Id.* at 123–25. There is also an undercurrent in the story about Byerley being potentially deployed as a corporate agent, but, importantly, there is no intimation that he would act merely at the corporation's behest. *Id.* at 116–17. Thus, there is no hint in the story that Byerley, although robotic, would not be autonomous—acting at his own behest toward his own, self-defined ends.

separate consciousness, is distinct in the way that consciousness features as *emergent* rather than extant and *alien* rather than recognizable. Although Byerley is functionally equivalent to a human (some would say superior to one, because bound by the Three Laws), the question about his selfhood—about what kind of *there* is there—revolves around the notions that (1) his being has its basis in human creation—if made, he was made by some human entity, likely a corporation—and, (2) if a robot, his way of being in the world is of a different kind than that of humans. In this way, the story of Stephen Byerley reveals one of the fears at work in the Automaton encounter—namely, the way in which the appearance of a sentient machine would disrupt the categories currently used to make sense of the world. We're used to the category of consciousness tracking that of humanity, but in the imagined encounter with an Automaton, we see the connection between those two categories dismantled in a manner that forces us to question our intuitions about how to relate to the world. If human-ness is no longer a condition of operating in the world, we are suddenly forced to ask ourselves what *does* matter instead. This existential problem is not dissimilar to the one forced by the encounter with a Bad Ex-Boyfriend; indeed, it's very much the same problem, just presented in a different guise: what are the standards for capacity to operate as a being in the world? If the Bad Ex-Boyfriend seems to have failed to live up to them, prompting psychological horror and subsequent philosophical inquiry into the assuredness of his selfhood, the Automaton raises the issue of whether he is encompassed by them at all. Are such emergent beings, different in kind from what we are already familiar with, essentially capable of meeting these standards? Will they be able to apply and abide by them?

It is perhaps unsurprising, then, to realize that in both instances the ensuing philosophical conversation often takes on the language of emotion. While we don't often think of emotion as necessarily central to selfhood, we do tend to treat it as somehow instrumental in guiding appropriate interaction in the world. I am compelled to describe my Bad Ex-Boyfriend's selfishness in terms that highlight his inability to feel or project emotions—he's heartless, or cold. In the same way, while we become increasingly accustomed to the idea of encountering intelligent Automatons—as we perhaps already arguably do in the case of algorithmic decision-making—we begin to couch our worries about the Automatons of the future in similar, emotion-laden terms. We worry, in metaphorical terms, about Automatons being heartless, an articulation of the problem that brings to the fore related questions about the relationship between emotion and consciousness and that between emotion and humanity. Does emotion accompany consciousness? Or do you need—at least as a necessary, if not a sufficient, condition—humanity for that?

Although this at first appears perhaps as a kind of artificial goal-post-movement (having encountered robots with emergent intelligence, humans, still fearing them, suddenly raise the standard for living amongst us to provable emotional capacity), the problem of emotion is actually intimately embedded in the concept of consciousness and/or sentience itself. To see this, we need look no further than the case of the Bad Ex-Boyfriend: the encounter with him raises the issue of selfhood by nature of the fact that it illuminates his emotional emptiness;

seeming to lack the requisite affective capacity, he exposes his *self* to existential skepticism—and the ensuing philosophical discussion asks whether his demonstrated emotional deficiency might in some way preclude him from the designated benefits that accrue in virtue of being categorized as "human" in our world. With the Automaton, the issue appears somewhat inverted (again, if only because it is difficult to escape the intuitive notion that automatons are, by their very nature, distinguishable from humans in some way):[73] it appears as one of endowment rather than degradation—namely, would an unfeeling autonomous robot merit selfhood?

The relevance of such a question has not been overlooked by computer scientists and AI technologists, and "[r]esearch on the role emotion plays in cognition has entered computation."[74] Although "[m]ostly these researchers are . . . working . . . to create better simulations of minds with emotion or affect,"[75] they are, according to Siri Hustvedt, plagued by the problem of distinguishing "felt emotions [from] the appearance of emotions."[76] As any scorned ex-girlfriend knows, this problem is a real one, and it can be illustrated well through an exploration of Cynthia Breazeal's "sociable robot" Kismet. Kismet is a "big-eyed interactive robot head" capable of simulating "infantile emotional facial responses."[77] Kismet can talk, and it can make "expressive facial movements and sounds that are pitched to mimic emotion in relation to [an] interlocutor."[78] Breazeal has described Kismet as having a "synthetic nervous system," and a "motivational system," as well as "six basic emotions," including anger, disgust, fear, happiness, sadness, and surprise.[79] The question we might next ask about Kismet, which Siri Hustvedt does ask, is the same one we might ask of a future Automaton: can this machine, which simulates emotion, and, perhaps more importantly, triggers emotional connection in humans, be called an *emotional* machine?[80] Part of the answer would seem to lie in the operative distinctions between computation and reasoning. Those who ascribe to a computational theory of mind (CMT) may be inclined toward the view that a machine could be emotional to the same extent as a human, or that it could be conscious independent of its ability to feel, while those who advocate for an embodied understanding of cognition might be both more skeptical that emotion could exist beyond its embodied, lived-in form and more convinced that emotion is an integral part of consciousness.[81] If emotion can't be computationalized, it's hard to imagine emotional machines. Whether or not this *matters* may have to do with the extent to which emotion is considered important for the sort of everyday reasoning

[73] *See supra* note 51.

[74] Hustvedt, *supra* note 1, at 268.

[75] *Id.*

[76] *Id.* at 273.

[77] *Id.* at 272.

[78] *Id.*

[79] *Id.* 272–73.

[80] *Id.* at 273–75.

[81] *See id.*

Deleted: 5151

that intuitively strikes us as necessary to operating well in the world. Siri Hustvedt articulates the matter thusly:

> In . . . artificial intelligence, feeling boredom, joy, fear, or irritation must be turned into a rational process that can be translated into symbols and then fed into the computer. Emotion has to be lifted out of a feeling bodily self. That is not easy to do. When I'm sad, can that feeling be parsed purely through logic? One may also ask if it is possible to *reason* well in our everyday lives without feelings. It is now widely acknowledged that emotion plays an important role in human reasoning. Without feeling, we aren't good at understanding what is at stake in our lives. And therefore psychopaths and some frontal lobe patients who lose the ability to feel much for others are profoundly handicapped, despite the fact that a number of them can pass tests that show they have no "cognitive" impairment and can "compute" just fine. They may well be able to follow the sequence of a logical argument, for example, but they suffer from an imaginative emotional deficit, which results in their inability to plan for the future and protect themselves and others accordingly. And if this affective imagination is not a conscious act . . . how can it be programmed into a machine?[82]

In this sense, the problem of encountering of an Automaton consciousness may lie in the way that this consciousness, emerging from a machine, is alien. Because of this alien-ness, in the case of the Automaton, we lack a certain assuredness that this autonomous being, though capable of interacting on our level and of convincingly simulating much of our behavior and physicality—of seeming, in other words, to have a *there* there—will be the same *kind* of thing that we are. The psychological horror that accompanies encountering the Automaton thus takes the shape of our intuition that an entity needs to be the same kind of thing as we are in order to be capable of selfhood—in order to have the potential to feel as we do and, perhaps, reason as we do. The ensuing philosophical discussion, then, articulates the problem of the self in terms of consciousness, including both its constitution and its sufficiency as a standard-bearer for operating in the world.

2. The Philosophical

The philosophical quandaries raised by possible future encounter with seemingly-sentient Automatons involve questions about the concept of the self, but, with this scenario, the concept of the self is articulated in terms of humanity rather than terms of consciousness. Because Automaton consciousness is, by definition, alien—emanating from a source that we would not intuitively expect—the horror of this confrontation is wrapped up in an effort to try to apply our existing categories to determine whether such consciousness is equivalent to ours

[82] *Id.* at 266–67.

in the ways that are important for living together. If the Automatons of the future strike us as real, and, like Stephen Byerley, seem almost as if they could be people, is there any reason not to treat them—at least in public life—as such? In considering whether these Automatons have selves, we are in large part asking the following: is there any operative ontological characteristic that prevents them from capably meeting the standards of shared life? Thus, while with the Bad Ex-Boyfriend selfhood became a language for negotiating shared standards of how a person should be, and we were able to see the way in which our meaning-making categories elided into one another and stood in for something else, with the Automaton we face not only these difficulties but another, more foundational one: the figure of the Automaton, seemingly conscious but certainly not human, capable of being almost exactly like us, save for the inchoate, dream-like doubt he triggers, threatens to explode our categories altogether.

To see how this is the case, it is helpful to compare our future Automaton encounter with an often-invoked philosophical test case—that of the philosophical zombie.[83] The philosophical zombie problem posits a being who, like our future automaton, "is indistinguishable from a normal human being."[84] These zombies, again like our Automaton, are "capable of acting in all the ways in which a typical human being can act."[85] They can "move around . . . solve problems in creative and imaginative ways, tell jokes, learn new skills and language, appreciate music and art, write novels, be loyal (or disloyal) friends."[86] They also "talk about all the things a normal human talks about."[87] But they don't, however, actually have experiences. There is, when it comes right down to it, no *there* there. Philosophical zombies are "devoid of conscious thought and sensory experience of any kind;"[88] although they claim to have these experiences—a "zombie will say that it is in pain if it breaks a leg, but in reality it feels nothing"[89]—they do not actually have them. They may have, as philosopher Barry Dainton puts it, a "distinctive kind of mind,"[90] different from ours but capable of delivering functionally equivalent outputs. However, because there is not actually any *there* there, "there are good reasons for regarding zombie selves—and lives—as possessing less intrinsic worth than normal conscious selves."[91]

This comparison to the zombie problem illuminates the two strains to our Automaton philosophical inquiry. In the first place, if our future Automatons might be like these zombies—distinctive kinds of beings, but familiar enough that we are prompted to inquire after their selfhood—we may be driven to ask ourselves whether they, like the Bad Ex-Boyfriend, are entitled to a benefit of the

[83] *See supra* note 13 and accompanying text; *see also* DAINTON, *supra* note 3, at 190–96.

[84] DAINTON, *supra* note 3, at 191.

[85] *Id.* at 194.

[86] *Id.*

[87] *Id.* at 191.

[88] *Id.*

[89] *Id.* at 194.

[90] *Id.* (emphasis omitted).

[91] *Id.* at 194–95.

doubt.[92] If we project their existence as arising from the continued development of AI technology, we might eventually encounter Automatons who expose the conceptual fault line problematized above by Siri Hustvedt using the touchstone of emotion: the difference between simulating and being, between expressing emotion and feeling it, between acting as if one has experiences and actually having them.[93] Such future automatons might, borrowing Barry Dainton's depiction of philosophical zombies, "claim to have feelings and sensations—along with conscious thoughts and memories—but . . . really have [only] non-conscious states in their information-processing systems that they *call* 'feelings', 'thoughts', 'memories', and so on."[94] Because such beings might truly be lacking some important working piece—namely, ipseity—there would be, with them just as with philosophical zombies, "good reasons for regarding . . . [them] as possessing less intrinsic worth than normal conscious selves."[95]

The problem, of course, is just "how much less,"[96] as well as how to formulate this distinction, let alone enforce it. And this is where the efficacy of our categories starts to break down. How would we know which automatons had it and which didn't? If it's conceivably possible that future Automatons might have consciousness, but also possible that they might be more like philosophical zombies, couldn't there be some of one and some of the other? After all, the philosophical zombie problem is not just directed at projected machine intelligence; it also illuminates the fundamental epistemological problem reflected in the Bad Ex-Boyfriend analogy—the problem of other minds.[97] Under the weight of all these questions, it becomes difficult to see how selfhood (here articulated in terms of consciousness), although it seems intuitively important to us, can actually do the work we want it to. Thus, our asserted aim—of protecting the worth of that thing, the *there* there that we actually have, but that the evil fake only imitate—seems to slip away, and it looks like we are, again, having little more than a very complicated conversation about how beings (not just humans this time) should be.

[92] I do not, by this, assert that the Bad Ex-Boyfriend in fact *is* entitled to a benefit of the doubt where the Automaton is not; instead, I suggest that he is intuitively afforded one, perhaps as a function of the way we use our categories—i.e., perhaps we suspect that being human entails having consciousness, even if we harbor suspicions about the Bad Ex-Boyfriend's selfhood in light of Part III of the confrontation with him.

[93] HUSTVEDT, *supra* note 1, at 276. Hustvedt draws this comparison by pointing to the work of David Galernter, professor of computer science at Yale University. Galernter does not believe that "an 'intelligent' computer will ever *experience* anything. It will never be conscious. 'It will say . . . that makes me happy,' but it won't feel happy. Still: it will act as if it did." *Id.*

[94] DAINTON, *supra* note 3, at 194.

[95] *Id.* at 195.

[96] *Id.*

[97] The other minds problem deals with "how to justify the almost universal belief that others have minds very like our own," thereby "hav[ing] inner lives" that are the same kind of thing we have. *Other Minds*, STAN. ENCYCLOPEDIA PHIL., https://plato.stanford.edu/entries/other-minds/ (last revised Jan. 14, 2014).

This problem is the second strain of the Automaton philosophical inquiry. If encountering the Automaton threatens to degrade our concept of consciousness, what will we be left with? Will we be driven back to the shorthand of humanity, finding ourselves pouring new meaning into the category of having been provably born of a woman? Or will this category, too, start to feel arbitrary in light of the real, functional similarity between human and automaton? As Barry Dainton puts it, "[i]f non-conscious zombies can do everything we can do . . . it will look very much as though consciousness per se contributes little that is worthwhile or distinctive. . . . If so, can it really be right to value it very highly?"[98] The encounter with a future Automaton thus makes us question the sensibility of some of our most foundational meaning-making categories, forcing us to consider the ways in which they represent little more than content pieces in an on-going negotiation of the standards for being and for shared life. The confrontation of this philosophical possibility—that maybe the emperor really isn't wearing any clothes, that our deployment of concepts like selfhood is really a negotiation of norms all the way down—accentuates a reality that is too often overlooked in legal circles, which are apt to build legal rules upon philosophical concepts as if those concepts were ahistorical. This reality is that culture matters. Culture feeds and shapes the way we use indeterminate concepts to make meaning in the form of provisional solutions. Our provisional solutions to age-old problems that can't be definitively settled often come in the form of legal rules—particularly when those problems have to do with the requirements of living together in shared society. Thus, we would do well, in our on-going discussion of robot rights, and our attempts to prepare for the challenges of tomorrow's technology, to consider our prevailing discursive forms and the narratives that we're using to make sense of ourselves and others. In the second decade of the twenty-first-century, one of those narratives deploys the concept of the self in a manner that could prove impactful for the Automatons of the future.

II. Meaning Making—Culture, The Narcissism Script, and Imagining the Shared Society of Tomorrow

The Bad Ex-Boyfriend/Automaton analogy suggests that both figures may prompt similar psychological responses and lead to similar philosophical quandaries. It also provides insight into the ways in which our conceptual categories—like selfhood—function as battlegrounds for negotiation about the way we think people should be and the kind of society we want to have. Understanding that these indeterminate philosophical concepts may inform the way we think about important questions of acceptable conduct and capacity for participation in communal life demonstrates the necessity of examining the way such concepts are deployed, defined, and discussed in the culture at large. The narratives we reach for to explain ourselves when we encounter a Bad Ex-Boyfriend today or an Automaton tomorrow will reflect the discursive influence of cultural trends upon important but ultimately indeterminate concepts like the self. Thus, the prevalent narratives espoused around the self in the second decade

[98] DAINTON, *supra* note 3, at 195.

of the twenty-first-century may well have reverberating effects upon the laws we will be able to imagine as suitable for the possible automatons of tomorrow. With this in mind, let's explore one of the most common selfhood narratives: the narcissism script, which increasingly provides the cultural tools to diagnose a broad range of humans as self-less shells. After examining the contours of the script, we will ask what the prevalence of such a narrative might mean for the way we will treat future automatons.

A. The Narcissism Epidemic

It need not be the case that the psychological terror provoked by the emergence (or reemergence) of the other triggers a diagnosis of narcissism and an accompanying discourse denying selfhood. Instead, this is a cultural script grounded in social occurrences and historical developments. It is socioculturally bounded. In *The Selfishness of Others*, Dombek deftly captures the reality and nuances of what she calls "the narcissism script,"[99] wherein she identifies the defining aspects of this cultural narrative. In order to draw the contours of the narcissism script and demonstrate what it reveals about the dynamic of collective conceptions of selfhood in the early-twenty-first century, I will explore three of the script's defining features.

To begin with, the narcissism script invokes what it claims to be a true rise in the incidence of Narcissistic Personality Disorder (NPD).[100] This rise in clinical cases of narcissism is simultaneously caused by historical factors—like the pervasion of social media and the particularly over-coddling parenting trends of the late-twentieth century[101]—and accompanied by an increased pop cultural interest in the disorder that provokes expanded usage of the term as a metaphorical descriptor. For instance, Neil J. Lavender, a professor of psychology at Ocean County College and a *Psychology Today* blogger, states that "there are more narcissists today living in the United States than at any other time, with the millennial generation leading the pack,"[102] a view he attributes to the broader community of "mental health professionals."[103] Jean Twenge and W. Keith Campbell, co-authors of the 2009 book *The Narcissism Epidemic*, agree with him.[104] Their seminal work reports the results of a study showing that "millennials [test] higher on the Narcissistic Personality Inventory than any generation before," and that "[o]ne in ten Americans in their twenties . . . 'has experienced symptoms' of' NPD."[105] Other works, such as *Narcissists Exposed, Why Is It Always About You?* and *The Narcissist Next Door* report similar statistics.[106] Picking up on these findings, "bloggers, journalists, and pundits" have taken the narcissism script

[99] Dombek, *supra* note 2, at 10.

[100] *See id.* at 19.

[101] *See id.* at 10.

[102] *Id.*

[103] *Id.*

[104] *See id.* at 18–19.

[105] *Id.* at 19 (quoting Jean M. Twenge & W. Keith Campbell, The Narcissism Epidemic: Living in the Age of Entitlement (2009)).

[106] Dombek, *supra* note 2, at 19.

mainstream, diagnosing a broad range of public figures as narcissists in an effort to demonstrate the moral degradation of society at large.[107] To borrow from the title of a 1979 book that espoused similar ideas about a prior generation, such public thinkers describe the way in which we now live within a "Culture of Narcissism,"[108] with "Generation Me" being emblematic of an "epidemic of contagious and toxic self-absorption,"[109] which can be seen manifested in everyone from Kanye West and the Kardashians to Oprah Winfrey, Eckhart Tolle, John Edwards, and Barack Obama.[110] According to *Huffington Post* blogger Ike Agwu, social media is at least partially to blame: "Twitter convinces [a generation of millennials] that they are being followed and are special," that what they do and "like" is automatically endowed with importance.[111] But other narcissism-laden cultural developments abound: in 2013, Oxford Dictionaries declared "selfie" to be the Word of the Year, and in the second decade of the twenty-first century "our language is more self-centered than ever before," with "American writers using *I* and *me* forty-two percent more than they did in 1960."[112] Such developments seem to suggest that NPD has "spread through the culture like a disease,"[113] and, indeed, judging by the metric of the Narcissistic Personality Inventory (NPI), a survey developed by social psychologists Robert Raskin and Calvin Hall in 1979 to measure traits of NPD,[114] the contagion may indeed be real. Psychologist Jean Twenge used this metric, collecting and analyzing 16,475 NPI surveys of college students, to report a thirty percent rise in narcissism from 1979 to 2006.[115] Though her results have been contested,[116] the cultural refrain persists. Although, "[s]ince 1980, when NPD was first introduced as a diagnosis . . . the American Psychiatric Association has claimed that less than 1 percent of the population suffers from it,"

[107] *Id.* at 20.

[108] *Id.* at 62; *see also* CHRISTOPHER LASCH, THE CULTURE OF NARCISSISM (1979).

[109] DOMBEK, *supra* note 2, at 62.

[110] *Id.* at 20.

[111] *Id.* at 20.

[112] *Word of the Year 2013*, OXFORD LIVING DICTIONARIES (2018), https://en.oxforddictionaries.com/word-of-the-year/word-of-the-year-2013; DOMBEK, *supra* note 2, at 11. One recent study, however, found that there is not a connection between first-person language (or "I-talk") and narcissism. *See* Angela L. Carey et al., *Narcissism and the Use of Personal Pronouns Revisited*, 109 J. PERSONALITY & SOC. PSYCH. e1 (2015).

[113] DOMBEK, *supra* note 2, at 19.

[114] Dombek notes that the NPI was designed to measure traits of NPD in "normal" and "healthy" people. *Id.* at 66. According to one study, "[t]he vast majority of research in social/personality psychology uses various forms of the Narcissistic Personality Inventory" to assess narcissism, despite there being "increasing concerns about the conceptual underpinnings and psychometric properties" of the measure. Robert A. Ackerman et al., *What Does the Narcissistic Personality Inventory Really Measure?*, 18 ASSESSMENT 67, 67 (2010), http://www.sakkyndig.com/psykologi/artvit/ackerman2013.pdf.

[115] DOMBEK, *supra* note 2, at 68.

[116] *Id.* at 68–76.

and has recently revised the figure to between zero to five percent, in the cultural imagination narcissistic bogeymen abound.[117]

The notion that our culture is increasingly narcissistic—both creating and being created by a rise in clinical narcissists—informs the second important aspect of the narcissism script: the exaggeration or egalitarianism of its application, whereby we now diagnose our bad ex-boyfriends with the same kind of "selfishness" once reserved for psychopathic serial killers.[118] Perhaps because "NPD is no longer markedly different from the expectations of our culture, but our culture exactly,"[119] there appear to now be as many varieties of narcissism as there are types of people. Dombek chronicles the seemingly endless list of various strains, and in so doing draws attention to the way in which this expansion of the category threatens to stretch it beyond comprehension:

> If you are 'reckless' and 'self-assured' and a 'social climber,' you are a Phallic Narcissist. If you are in a group that thinks its members are more special than other people in the world, your group has Collective Narcissism. If you are the leader of a group, company, or country and are motivated by grandiose ego rather than care for your constituents, you are a Narcissistic Leader. If you are in a culture that prioritizes superficial symbols of power such as wealth and all you care about is competing for said symbols, then you are participating in Cultural Narcissism. If you are the leader of a corporation and you have only one thing on your mind—profit—then you are a Corporate Narcissist. . . . If on the surface you think you are really spiritual and seek out religious structures and practices to confirm this . . . you are a Spiritual Narcissist. If you are a scientist whose sense of your own genius leads you to dominate dinner table conversations . . . you are a White Coat Narcissist . . .[120]

If, indeed, the narcissist is an "evil fake", and narcissism has really expanded in this way, "spreading like a virus through the thin, recirculated air of modern American culture,"[121] the problem does seem to be of epidemic proportions. If everyone from the social-media obsessed teen to the unsympathetic doctor to the oblivious boyfriend can be shown to manifest some form of narcissistically disordered behavior, it starts to look very much as if "we live in a time so rampant with narcissisms . . . that ours is a moment in history that is . . . absolutely exceptional."[122] The problem with this tendency to apply the same descriptor to such a broad range of disfavored behavioral traits is, of course, that it starts to rob the descriptor of any true meaning: if narcissism can be both everything and

[117] *Id.* at 18.

[118] *See id.* at 9.

[119] *Id.* at 18.

[120] *Id.* at 21–22.

[121] *Id.* at 19–20 (quoting Twenge & Campbell, *supra* note 105).

[122] *Id.* at 11.

nothing, can it really be anything at all? And if narcissism entails the absence of a self, aren't we now dealing with a massive number of "false selves" out there "masquerading as real selves"?[123]

The second aspect of the narcissism script—its expansiveness—is rendered even more problematic by the third and final one: the fact that the proffered solution to this posited problem closely mirrors the diagnosis itself. The best example of this aspect may be the advice offered to scorned, long-suffering ex-partners of selfish narcissists. Much literature—in the form of books, blogs, and online support communities—provides advice on how to know when you're in a relationship with a narcissist, how to parse his or her narcissistic behavior,[124] and what to do about it.[125] Dombek cites no less than seven websites designed to provide healthy selves with the tools necessary to break free from the empty narcissists in their lives.[126] The myriad of advice seems ultimately to resonate with a single command: leave now, don't look back, get out as fast as you can.[127] The internet support groups call this "going 'no contact,'" and one has even created an app to help weak victims actualize: the app prompts your smartphone to send you a message on command designed to renew your resolve to stay away.[128] The message reads, in part: "Do NOT call them! . . . No Contact is one of the most hurtful narcissistic injuries you could inflict."[129] The irony embedded in the message is not lost on Dombek, and should not be lost on us: while the message means to suggest that going "no contact" will be one of the most painful things you can do to the narcissist in your life—because narcissists feed on the attention of others[130]—it ambiguously describes the encouraged behavior itself as narcissistic.[131] This example cleanly epitomizes the way in which the solution offered for dealing with narcissists mirrors the problematic dynamic of narcissism itself: victims are encouraged to "enact[] the very coldness described by the diagnosis, as if the only way to escape the emptiness contagion is to act like a narcissist yourself, and turn away from anyone flat and fake."[132] In the end, then, this aspect reflects the very point that this Part is trying to make—namely, that narrative is important. The way we describe our culture can quickly become the

[123] *Id.* at 12.

[124] There's an entire sub-language for this, which provides terms that unveil the narcissistic motivation behind seemingly innocent behavior. For instance, a disinterested boyfriend might be "doing a discard" and a family member keeping sentimental mementoes might be "keeping a trophy box." *Id.* at 25–27.

[125] *Id.* at 23.

[126] *Id.* at 23.

[127] *See id.* at 10, 25.

[128] *Id.* at 25.

[129] *Id.*

[130] *Id.* According to the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* (5th ed. 2013), one of the diagnostic criteria for narcissism is "requir[ing] excessive admiration." *See* Leon F. Seltzer, *6 Signs of Narcissism You May Not Know About*, PSYCHOLOGY TODAY (Nov. 7, 2013), https://www.psychologytoday.com/blog/evolution-the-self/201311/6-signs-narcissism-you-may-not-know-about.

[131] DOMBEK, *supra* note 2, at 25.

[132] *Id.* at 12.

way it is, and this is nowhere clearer than in the metastasis of the narcissism script. At first, it highlights just a few more clinical narcissists; soon, whole swaths of individuals become, if not clinically narcissistic, at least close enough; finally, those spared few become, in their effort to fight against the narcissists at the gates, little better than narcissists themselves. As Dombek puts it:

> The script confirms itself, and the diagnosis and treatment confound the evidence, until it gets harder and harder to know whether people are really more selfish than ever before in the first place. In this way, it matters whether or not it's actually real, but it matters even more whether or not we *believe* it's real.[133]

In the narcissism script, then, we have a veritable crisis of selfhood. If narcissists are selfless, empty entities, and they are taking over the country—and maybe even the world[134]—it's starting to look like there might be fewer and fewer selves out there. Indeed, it's starting to look like there might not really be any at all. In this way, the narcissism script enacts an instantiation of the fundamental skepticism surrounding selfhood. In the cultural ascendance of this narrative, we can see the echo of both our fundamental doubts about ipseity and the way in which the concept of the self becomes a category we use to engage with and critique our modes of living together.[135] But if we are now living in a world increasingly populated by a bunch of empty, self-less humans, what might that presage about our future treatment of automatons?

B. Automatons, They're Just Like Us!

The foundational skepticism about ipseity animating the narcissism script also characterizes much of modern AI research and theory. Though their work seems to implicate fundamental questions about the nature of consciousness and selfhood, many of the leading thinkers and researchers in this area have simply put this question to one side, content to focus on more concrete computing and engineering problems.[136] Either that or they espouse the view that the existence and nature of the self is at best beside the point—that there is, in the end, no meaningful distinction between simulating selfhood and having it. A good example

[133] *Id.*

[134] *Id.* at 10. As one blog exemplifying this notion put it, "[t]he term 'narcissist' seems to be spreading through the world like an out of control wildfire." Tina Swithin, *The Narcissism Epidemic: Do You Know the Warning Signs?*, HUFFPOST BLOG (Aug. 27, 2013), https://www.huffingtonpost.com/tina-swithin/the-narcissism-epidemic-k_b_3510595.html.

[135] Such a notion is perhaps strengthened by the insight that Millenials aren't the first generation to be diagnosed with epidemic narcissism. Indeed, the narrative has been deployed at other times throughout history, particularly in the mid-twentieth century. *See* DOMBEK, *supra* note 2, at 62–63.

[136] The ability of such thinkers to put this problem to one side is interesting when considered in light of the fact that many of them adopt a theory of mind that minimizes the difference between humans and machines by conceiving of human intelligence as a kind of computing and of humans themselves as a kind of machine (at least as regards the inner workings of the mind and/or brain). *See* HUSTVEDT, *supra* note 1, at 254–63.

of this attitude comes from Rodney Brooks, Panasonic Professor of Robotics at MIT and former Director of the MIT AI Lab, who, in conversation with author and computer programmer Ellen Ullman, admitted to "view[ing] the interior life rather cynically, as a game, a bunch of foolery designed to elicit a response."[137] In her work *Life in Code*, Ullman recounts a conversation with him about the self: "I asked him, 'Are we just a set of tricks?' He answered immediately. 'I think so. I think you're a bunch of tricks and I'm just a bunch of tricks."[138] Brooks, it seems, is unpersuaded that there is something fundamental happening inside of human beings that makes them meaningfully different from advanced computational machines. Indeed, in his book *Flesh and Machines: How Robots Will Changes Us*, he writes that "[a]nything . . . living is a machine. I'm a machine; my children are machines. I can step back and see them as being a bag of skin full of biomolecules that are interacting according to some laws."[139] Moreover, although Hustvedt reads Brooks as being "well aware that his 'creatures' [robots] don't *feel* the way human beings do," she reports that he nevertheless finds the "difference between 'us' and 'them' [to be] insignificant."[140] Such views might seem eccentric or surprising until we remember the force and prevalence of the narcissism script: judging by that, Brooks's view of selfhood is very much the mainstream—though we may feel intuitively that humans are different from automatons, we'll have trouble tracking this difference to ipseity. After all, we are very fond of talking about how many people out there walking around right now are really just empty fakes. What, therefore, when faced with the Automatons of the future, could we say marks them out as meaningfully unlike us?

Such a question suggests the way in which the cultural discourses of today shape the imaginative capacities of tomorrow. Is the narcissism epidemic a crisis of selfhood or the revelation of it as conspiracy? By drafting so much of our critique of modern life in a narrative that denigrates people's selfhood, have we communicated that we value the concept or instead reified its irrelevance? The answers are myriad, likely contradictory, and illustrative more than anything else, but it's possible to envision the narcissism narrative affecting our potential future encounter with automatons in at least three sorts of ways. Under each possibility, a different kind of treatment would seem sensible, thereby influencing the legal structures that eventually appear intuitive and socially acceptable.

In Scenario One, the script could prove beneficial to the Automatons by encouraging the adoption of a new concept as the meaningful categorical designator of capacity for participation in communal life. With humanity off the table, that characteristic might be consciousness.[141] If selfhood is in doubt with both Bad Ex-Boyfriends and Automatons, but we no longer have humanity to fall back upon in the case of the Automatons, perhaps we will decide that consciousness is enough. The problem with this is that this concept, though less tainted than that of the self by the cultural discourse surrounding narcissism, is

[137] Ellen Ullman, Life in Code 156 (2017).
[138] *Id.*
[139] *See* Hustvedt, *supra* note 1, at 270.
[140] *Id.* at 271.
[141] *See supra* Part I.

equally fraught. Indeed, the two concepts might actually be the same thing, plus or minus an implication of humanity.[142] The prevalence of the narcissism script, may, in combination with the alien-ness of the automatons, make it seem insensible to hold a conversation about Automaton selfhood, thus driving the negotiation of Automaton treatment to a working out of what qualifies as consciousness. If, as some technologists suggest, there's no meaningful difference between seeming and being—simulating and experiencing—and if humans are, after all, very much like machines—with computation at the core of intelligence—perhaps consciousness might eventually be quantified as a certain level of computational complexity. If this comes to be, maybe we can, in our future confrontation with Automatons, leave aside all of the thorny problems about feeling and interiority (leaving the question of seeming versus being to the realm of horror films) and get on with life alongside humanoid robots. In this case, perhaps the erosion of selfhood accomplished by the narcissism epidemic might prove beneficial for future automatons, driving humans to treat other conscious beings on equal terms, as capable of living with us in communal life. Maybe the automatons will have the Bad Ex-Boyfriends of yesterday to thank for shifting our operative conceptual category from selfhood to consciousness—a concept with which sense can be made of them.

Scenarios Two and Three, however, suggest less positive outcomes for the Automatons, one creating the possibility for a sort of passive discrimination, and the other for a stronger form of prejudice. In Scenario Two, the cultural prevalence of the narcissism script might spell circumscribed or somehow differential treatment for future automatons that derives from a skepticism about their selfhood buffeted by appeals to their alien-ness. If we understand the narcissism epidemic's use of the concept of selfhood as a degradation of the self, which denies ipseity to Bad Ex-Boyfriends and anyone else who has failed to behave in a manner deemed appropriate,[143] we might find it quite easy to translate this narrative to Automatons, who are perhaps more likely, given their machine provenance, to be one of the evil fake than other humans.

Support for such a reading might be found in the way that the denial of selfhood to narcissists often functions as a challenge to their *humanity*—in the sense that we mean, by denying these individuals' selfhood, that there is something fundamentally wrong with their way of being in the world.[144] Such a reading might not be altogether wrong-headed, especially if we believe Siri Hustvedt's suggestion that emotion plays an integral part in reasoning well. If that is the case, where the narcissism script helps to identify true incapacity to feel this sort of emotion, as in the case of psychopaths or others suffering from frontal lobe impairment[145], it may indeed tip us off to those—including future automatons—

[142] *See supra* note 7.

[143] This objective "appropriateness" often elides into "insufficiently attentive to *us*," a point Dombek makes subtly and evocatively in her essay. DOMBEK, *supra* note 2, at 45–59.

[144] *See supra* Part I.

[145] Such injuries have been shown to be capable of impairing the capacity to feel empathy. *See* HUSTVEDT, *supra* note 1, at 266–67; *see also* Arielle de Sousa et al.,

who lack something necessary to living within human community.[146] In other words, any human, or other being, who really *cannot* feel for others might well qualify for differential treatment, perhaps in the form of circumscribed autonomy. The complicating aspect, of course, is that it's not always the demonstratively impaired capacity for fellow feeling that we point to with our narrative of narcissism; instead, the narrative has allowed the term to metastasize such that we now apply it to anything we don't really like. And, although to have a personality disorder is, by definition, to fail to behave in a way that lives up to the "expectations of [one's] culture,"[147] that aspect of the narcissism script that expands the descriptor almost beyond sensibility reminds us that the cultural goalposts are moving and that where they land in any individual case is ultimately in the hands of a few decision-makers. After all, in a world where everyone is acting like a narcissist, but some are doing it only because they've been victimized by the *real* narcissists, it's easy to see how a category that could be (or once was) concrete can quickly become slippery and ephemeral. While the cultural degradation of the self via the narcissism epidemic hasn't yet been translated into legal structures—we've not yet seen a widespread movement to take away the rights, in public life, of narcissists—it's possible to imagine that with the subtraction of humanity from the equation, this extra step might be easily taken. Sure, we might suspect that a lot of humans out there don't have selves, but they're sufficiently like us to merit the benefit of the doubt. They get equal treatment in public life *until* their behavior demonstrates an impaired capacity, at which point they may find their rights circumscribed.[148] Automatons, on the other hand, lacking both a self *and* humanity, may seem distinct enough from us, intuitively, to merit differential treatment from the outset. If such treatment were to be instantiated in law, it might amount to a kind of discrimination.

Finally, in Scenario Three, Automatons could meet with a stronger sort of differential treatment than the one described above. Scenario Two's type of differential treatment would be driven by the intuitive sense that automatons are just *too* different because they both (1) might lack selves and (2) are not human. It's a sort of passive application of an intuition that combines (1) the selfhood notion, as emptied out by the narcissism epidemic, with (2) the vague sense of affinity represented by humanity to result in skepticism about treating automatons as capable of full participation, on equal terms, in human communal life. But, it's possible either that the narcissism epidemic gives way to a more complete dissolution of the concept of selfhood, and/or that consciousness proves too nebulous to be workable, a reality that might push society toward a strong, active reliance on the concept of "human-ness," such that capacity for participation becomes dependent upon provable human provenance, with differential

Understanding Deficits in Empathy After Traumatic Brain Injury: The Role of Affective Responsivity, 47 CORTEX 526 (2011).

[146] *See* HUSTVEDT, *supra* note 1, at 266.

[147] DOMBEK, *supra* note 2, at 17.

[148] For example, those with personality disorders may be subject to involuntary civil commitment. *See generally* Megan Testa & Sarah G. West, *Civil Commitment in the United States*, 7 PSYCHIATRY 30 (2010).

treatment meted out to all those beings who don't meet what might be, in the future, an increasingly arbitrary concept.[149] Such a future scenario would be bad for the Automatons, who, though in many ways analogous to Bad Ex-Boyfriends, might find themselves subject to differential treatment and exclusion from certain aspects of public communal life purely on the grounds of their alien-ness. This instantiation of cultural mores might amount to prejudice.[150]

III. Thought Experiment: Automatons Amongst Us

Fast forward a few decades—or, if you like, simply imagine a world very much like ours, but which also features the presence of a certain sort of human-esque, Turning-Test-passing, seemingly-sentient automaton. AI technology has developed to the extent that such automatons have become somewhat commonplace. Maybe they have taken over a variety of previously-human tasks, at which they have proven better and more reliable than us—like performing surgery or operating vehicles or policing the streets.[151] Now imagine that those automatons are subject to a different set of laws than humans. Such differential treatment might entail two things: (1) the circumscription of certain freedoms—like freedom of movement, with automatons being subject to rules that prevent them from going to certain places; and (2) the refusal to extend certain rights of citizenship to automatons—as in, for instance, Asimov's imagined *I, Robot* universe, where robots lack both the right to privacy and the right to own property.[152] Would the differential treatment of automatons in this imagined, projected future amount to wrongful discrimination? I hope that this thought

[149] Consider, for instance, some of the following questions that arise: What would be the bottom line in qualifying as human? Would it be being born? What if you'd been born but then uploaded yourself into a computer or become embodied in a different kind of machine, thus leaving behind your original physical form? What kind of record keeping might be entailed if this were the criterion of "human-ness" in a future where technology makes such things possible? *See generally* Dainton, *supra* note 3, at 176–204. For a representative incarnation of this problem from popular culture, consider the Voigt-Kampff test from *Blade Runner* (1982). *See generally* Lorraine Boissoneault, *Are Blade Runner's Replicant's "Human"? Descartes and Locke Have Some Thoughts*, Smithsonian.com (Oct. 3, 2017), https://www.smithsonianmag.com/arts-culture/are-blade-runners-replicants-human-descartes-and-locke-have-some-thoughts-180965097/.

[150] This imagined scenario would also have reverberating impacts on some other legal concepts—most glaringly, the legal fiction which treats corporations as people for some purposes. *See* John Niman, *In Support of Creating a Legal Definition of Personhood*, 3 J.L. & Soc. Deviance 142, 161–66 (2012).

[151] Robots have already been deployed in public safety capacities on the streets of Dubai. *Police Officer Goes on Duty in Dubai*, BBC News (May 24, 2017), http://www.bbc.com/news/technology-40026940. Robots are also already used in surgery. Eliza Strickland, *Autonomous Robot Surgeon Bests Humans in World First*, IEEE spectrum (May 4, 2016), https://spectrum.ieee.org/the-human-os/robotics/medical-robots/autonomous-robot-surgeon-bests-human-surgeons-in-world-first. Self-driving cars are, of course, also on the horizon. Kyree Leary, *After Early Tests, Google is Focused on Fully Self-Driving Cars*, Futurism (Oct. 31, 2017), https://futurism.com/after-early-tests-google-is-focused-on-fully-self-driving-cars/.

[152] Asimov, *supra* note 65, at 125.

experiment will help to concretize the insights of Parts I and II by illuminating the ways in which discursive trends mediate the translation of philosophical concepts—like selfhood—into agreed-upon standards of conduct—i.e., legal structures. Whether such treatment is conceived of as discriminatory, and thus whether it strikes the social consciousness at large as reasonable, will be influenced by the conceptual tools we use to make sense of these imagined beings.

A. Discrimination: Counting Reasons, Expressing Animus, and Determining Moral Worth

Although this Part is not meant to be a comprehensive exploration of either the theory or practice of discrimination law,[153] I will examine two different articulations of anti-discrimination regulation's animating principles in order to explore their potential application to the Automaton scenarios.

There is both disagreement about what discrimination *is* as well as generally-accepted agreement that it comes in different kinds.[154] For the purposes of this Article, however, I will adopt a provisional definition that will allow us to explore whether discrimination might exist in relation to future automatons, and will limit the exploration to direct discrimination, thereby assuming that future differential treatment of automatons would be explicitly addressed to automatons *qua* automaton.[155] This still leaves us with quite a bit to explore.

Broadly speaking, the general concept of direct wrongful discrimination involves (1) differential treatment of an individual (2) who is a member of a "socially salient group,"[156] which (3) is to that individual's disadvantage, and (4) is explained by that individual's being (or being believed to be) part of the "socially salient group."[157] In other words, a central case of discrimination involves an agent meting out differential treatment "with the aim of imposing a disadvantage on persons for being members of some salient social group."[158] Moreover, in the legally relevant sense, wrongful discrimination is typically delimited to actions taken in the public sphere, meaning that although it is possible to discriminate by

[153] The field of scholarship is relatively expansive, and an attempt of this kind would merit its own substantial inquiry. For an introduction, consider: DEBORAH HELLMAN & SOPHIA MOREAU, EDS., PHILOSOPHICAL FOUNDATIONS OF DISCRIMINATION LAW (2013); TARUNABH KHAITAN, A THEORY OF DISCRIMINATION LAW (2015); KASPER LIPPERT-RASMUSSEN, BORN FREE & EQUAL? A PHILOSOPHICAL INQUIRY INTO THE NATURE OF DISCRIMINATION (2014).

[154] *See* HELLMAN & MOREAU, *supra* note 153, at 1–2.

[155] *See Discrimination*, STAN. ENCYCLOPEDIA PHIL., https://plato.stanford.edu/entries/discrimination/ (last revised Aug. 30, 2015).

[156] The notion of a "socially salient group" derives from Lippert-Rasmussen, and is meant to connote the idea that the group membership involved in discrimination must be of a kind "important to the structure of social interaction[] across a wide range of social contexts." Kasper Lippert-Rasmussen, *The Badness of Discrimination*, 9 ETHICAL THEORY & MORAL PRAC. 167, 169 (2006).

[157] *See* Kasper Lippert-Rasmussen, *Private Discrimination: A Prioritarian, Desert-Accomodating Account*, 43 SAN DIEGO L. REV. 817, 820 (2006).

[158] *Discrimination*, *supra* note 155.

choosing not to "befriend, marry, or be buried in the same graveyard[]"[159] (to give just a few examples) as another person, we typically take a "less censorious view" of this kind of behavior.[160] Thus, wrongful discrimination involves either: (1) a particular actor—namely, the state itself or any agent of it; or (2) an actor making decisions of a particularly public nature—of a kind that impact certain core areas necessary to operation in the world, such as housing, employment, or public accommodation.[161] Thus, the wrongfulness of discrimination relates to both its content and its form. First, there is "the idea[] that all members of a society ought to enjoy equality of opportunity," and then there is the accompanying notion that such equality "requires [both] that the state should not discriminate against anyone" and "that private individuals acting in a public capacity . . . should also not engage in such discrimination."[162] Restated another way, anti-discrimination regulation is animated by the assertion that some reasons for acting to an individual's disadvantage are impermissible in public life.[163] They are not allowed to "count." Thus, acting to a person's disadvantage simply because he or she belongs (or seems to belong) to a particular kind of group—for instance, "turn[ing] people down for jobs because they are women, or refus[ing] to rent flats to people because they are black"[164]—is wrongful. Why, exactly, is this wrongful? Well, there's the rub. There are a number of theories, of which I will articulate two.

First is the theory which relates the wrongfulness of discrimination to its connection with prejudice. According to this theory, discrimination is wrongful because it generates disadvantage from a "special vulnerability to prejudice or hostility or stereotype" and the "consequent diminished standing" such animus entails.[165] This theory locates the wrongfulness of discrimination in the reasons by which it is motivated.[166] Ronald Dworkin articulates this theory as the idea that "it is unacceptable to count prejudice as among the interests or preferences government [or actors in the public realm] should seek to satisfy."[167] Thus, differential treatment or "discriminatory legislation" will be "unjust" where no

[159] Lippert-Rasmussen, *supra* note 157, at 818.

[160] *Id.*

[161] This description is meant to broadly track the reach of the Civil Rights Act, 42 U.S.C. 21.

[162] Richard Arneson describes the reach as "roughly . . . the market economy." Richard Arneson, "Discrimination, Disparate Impact, and Theories of Justice," *in* Hellman & Moreau, Philoshopical Foundations of Discrimination Law 87–111 (2013).

[163] Including the modifier "disadvantage" here marks out the problem of advantageous discrimination as beyond the bounds of this discussion. It is nevertheless an interesting problem and oft-examined in the literature. *See generally* Larry Alexander, *What Makes Wrongful Discrimination Wrong? Biases, Preferences, Stereotypes, and Proxies*, 141 U. Pa. L. Rev. 149 (1992).

[164] John Gardner, *On the Ground of Her Sex(uality)*, 18 Oxford J. Legal Stud. 167, 167 (1998).

[165] Ronald Dworkin, *Affirmative Action: Is It Fair?*, 28 J. Blacks Higher Ed. 79, 80 (2000).

[166] *See* Ronald Dworkin, A Matter of Principle 66 (1985).

[167] *Id.*

"prejudice-free justification is available," or, at the very least, where "we cannot be satisfied" that the agent taking the action is "relying on a prejudice-free justification." [168] Under this account, then, anti-discrimination law protects against actions that are taken for certain reasons or motives.[169] Animus of the kind embodied in prejudice is, furthermore, a prohibited reason, and this may be on account of the denial of "equal concern and respect" that is encompassed by such animus,[170] or because of the irrationality that characterizes it.[171]

According to the account of discrimination that locates its wrongfulness in prejudice, that wrongfulness has to do with the *reason* for the action, which reason is either (1) constitutively immoral in that it denies the equal standing of individuals, or (2) fundamentally irrational in that it takes as a reason for acting in a certain way something that is not actually a reason for doing so.[172] If we formulate the project of exploring anti-discrimination law as one of identifying which reasons for action are allowed to "count" in public life, it's possible to conclude that the theory connected to prejudice gives us two principles: first, that irrational reasons may never count[173]; second, that where the denial of the "equal moral worth"[174] of individuals is constitutive of a reason, that reason may not count.

A second theory of the wrongfulness of discrimination is also premised on the notion of "equal moral worth,"[175] but it expands the application of this notion to account for actions which have the effect of denigrating such worth without necessarily requiring that its denial be built into the decision-making logic behind the discriminatory action. Under this theory, wrongful discrimination need not express the judgment that some individuals are of "lesser moral worth"[176]—need not arise from animus or prejudice—in order to diminish the principle of equal moral worth by impacting other aspects of an individual's life in ways that obstruct the realization of the principle on the ground.[177] In other words, while the "'equal moral status' principle . . . says that each is entitled to 'respect', to 'treatment with dignity' . . . or 'to be treated as someone who matters," this "nebulous, highly abstract, ideal" must be "given . . . content" by other values that "tell[] us how people must be treated."[178] Efforts to provide this content have taken various forms. Denise Réaume invokes Bernard Williams's contribution to the theoretical tradition, with his identification of "two aspects of personhood that are bound up with the idea of equal moral status."[179] According to Williams, equal moral status

[168] *Id.*; *see also* JOHN HART ELY, DEMOCRACY AND DISTRUST: A THEORY OF JUDICIAL REVIEW 153 (1980).

[169] Dworkin, *supra* note 165, at 80; *see also* DWORKIN, *supra* note 166, at 66.

[170] Dworkin, *supra* note 165, at 80.

[171] Gardner, *supra* note 164, at 168.

[172] *Id.*

[173] *See id.*

[174] Denise Réaume, *Dignity, Equality, and Comparison*, *in* HELLMAN & MOREAU, PHILOSOPHICAL FOUNDATIONS OF DISCRIMINATION LAW 8, 20 (2013).

[175] *Id.*

[176] Alexander, *supra* note 163, at 159.

[177] Réaume, *in* HELLMAN & MOREAU, *supra* note 174, at 20.

[178] *Id.*

[179] *Id.*

reflects the reality that individuals share "a range of basic material needs," as well as the notion that each is entitled to a level of "self-respect," which demands that persons be "abstracted from certain conspicuous structures of inequality" in the "ident[ification] and realiz[ation]" of their "own purposes."[180] Conceived in this way, discrimination might harbor a host of wrongfulness that does not derive from any kind of prejudice or animus—encompassing, for instance, exploitative treatment or "conditions that suppress or destroy consciousness."[181] Under this theory, discrimination's wrongfulness is not necessarily found in the reason motivating it, and thus it is possible that even rational reasons might be prohibited from "counting" in public life if they have the effect of undermining "equal moral worth" or its conditions precedent.

B. Evaluating the Scenarios: Differential Treatment, Human-ness, and the Benefit of the Doubt

Using these theories as a framework, let's now turn to an evaluation of Scenarios Two and Three above. As you will recall, Scenario Three involved a strong form of prejudice wherein the narcissism script's emptying out of the concept of selfhood helped to drive a focus on "human-ness" as the operative standard for capacity to participate in communal life. Scenario Two, on the other hand, involved a more passive kind of skepticism about automaton selfhood. When translated into legal realities, these cultural developments might take the following shapes. First, it's not difficult to imagine Scenario Three giving way to a regime of differential Automaton laws based on prejudice. A reliance upon the concept of "human-ness" as the category indicative of capacity to participate in communal life could easily give way to legal structures that instantiate differential treatment for automatons on the basis that such beings are of lesser moral worth. Thus, were our skepticism about selfhood to precipitate a shift to the concept of "human-ness," we might find ourselves building legal structures that discriminate against automatons out of a sort of speciesist animus—even where those beings, like in the imaginative case of Stephen Byerley, are virtually indistinguishable from the most decent of humans.

So much for the case of discrimination from prejudice; in our thought experiment, as in the modern world, that is the easy case. Scenario Two is trickier. Translated into a legal reality, we might imagine that this case could involve wrongful discrimination under the second theory, but it doesn't seem to involve prejudice. After all, in this case the cultural discourse doesn't encourage an explicit animus against automatons. Instead, it simply espouses a vague suspicion that they may not have selves and enables an unprincipled approach to the laws affecting them—an approach that is prone to keep the status quo tipped in favor of more familiar beings, even where those more familiar beings might be subject to a similar sort of skepticism. With this Scenario, then, we are left with, perhaps, some application of the equal moral worth theory. The questions we might ask ourselves, to discern whether this theory applies, are twofold. In the first place,

[180] *Id.*

[181] *Id.*

we might ask whether the suspicion that a being lacks a self constitutes a reason that should "count" in public life. In other words, is the vague sense that automatons are not quite enough like us, that they may lack some key piece necessary to functioning in communal life, sufficient to support differential treatment with regard to certain key rights like movement, owning property, and retaining privacy? If such beings were to appear conscious and sentient—virtually indistinguishable from humans in every external, relevant sense—would our vague sense combined with our conservatism provide reasonable grounds for discrimination in public life? Or, might this reason, particularly in light of the Bad Ex-Boyfriend analogy, strike us as pretextual, leaving us unsatisfied that it is, as Dworkin might say, entirely "prejudice-free"?[182]

An answer to this question might also help us approach another one, which asks whether future automatons would be entitled to the "equal moral concern" animating the second theory of anti-discrimination.[183] Parsing it in this way, the skepticism of Automaton selfhood can be understood as a doubt about whether such beings merit equal moral status. Thus, the question of differential treatment becomes the search for proof that Automatons are actually of lesser moral worth. Now, perhaps, if we could prove that they lacked interiority, or could demonstrate that despite being conscious, their form of consciousness lacked the sort of imaginative emotional capacity that Hustvedt takes to be central to effective reasoning, we would have proof of lesser moral worth and therefore a good reason for differential treatment. We would have a reason that is sensible in public life.[184] But without such proof, the mere suspicion, standing alone, is insufficient. This point can be developed further by accepting one thing and observing another.

First, let's accept that anti-discrimination law is designed to protect the equal moral worth of agents. Second, let's observe that the early-twenty-first-century prevalence of the narcissism script indiscriminately emptied out the concept of selfhood by invoking suspicion about its existence wherever individuals happened to do something that someone didn't like. Combining these two insights, we might suggest that the cultural discourse for which we may unconsciously reach when crafting the legal structures of tomorrow could denigrate selfhood such that it encourages us to deny the equal moral worth of the people around us who don't behave as we wish—and to do so, perhaps especially, when they enact even the most basic expressions of autonomy, like identifying and realizing their own purposes independently of what we might want them to do. Understood in this way, the narcissism script might have anaesthetized us to the denial of equal moral worth in a manner that we should be both aware of and careful about. Noticing

[182] DWORKIN, *supra* note 166, at 66.

[183] This, of course, is yet another way of asking what qualifies as capacity to participate in shared life, and demonstrates the way in which that negotiation can take on legal-theoretical, as well as philosophical, terminology. But can it tell us anything more about how to answer these two, interrelated, questions? Or does it instead demonstrate that debates about legal structures are yet another, ultimately more explicit, way in which we work out agreed-upon standards of living together?

[184] If such beings actually are of lesser moral worth, that would seem to be a rational reason to treat them accordingly.

this, however, we can return to the Bad Ex-Boyfriend/Automaton analogy to make a further point. If, in Scenario Two, we were to end up justifying differential treatment for Automatons on the basis of our suspicions about selfhood, but not extend such treatment to Bad Ex-Boyfriends, we'd have to search for a reason why. I'd suggest that such a reason might be found in our proclivity to give Ex-Boyfriends the benefit of the doubt. The idea, then, with other humans, is this: give even those who seem like empty fakes the benefit of the doubt—assume that they are of equal moral worth until it is proven otherwise.[185] The idea strikes me as sound enough to merit extension to the case of future automatons: because it is better not to deny the equal moral worth of another agent, the scales should always tip in favor of the benefit of the doubt. We should not risk violating another agent's equal moral worth on account of the limits of our own epistemology. Where equal moral worth is at stake, suspicion isn't a good enough reason for differential treatment. Thus, to subject imagined future automatons to differential laws would be discriminatory.

IV. The Discursive Translation Process

In the previous Part, I proposed that we imagine a world just slightly different from our own—a world where everything is basically the same, except that there are a bunch of life-like, humanoid, seemingly-sentient automatons wandering around. I then used this world to explore certain questions, questions both germane to doctrinal legal problems and revelatory of the discursive processes that often invisibly shape legal structures. Questions like: what might the laws in that world look life? How would humans and these almost-but-not-quite humans relate to one another in public life? And why? Unlike the short horror story from the very beginning of Part I (where your partner's eyes mysteriously turned from blue to brown overnight), this thought experiment wasn't meant to be a horror story. It was, however, meant to stir the sort of thoughtful examination often triggered by that kind of horror trope—namely, to show the cracks between the world as it is and the world as we make it. Lawyers are particularly fond of thinking that we are working with the world as it is; so much so that we are prone to overlook the ways in which the world is constructed and transformed through our involvement.

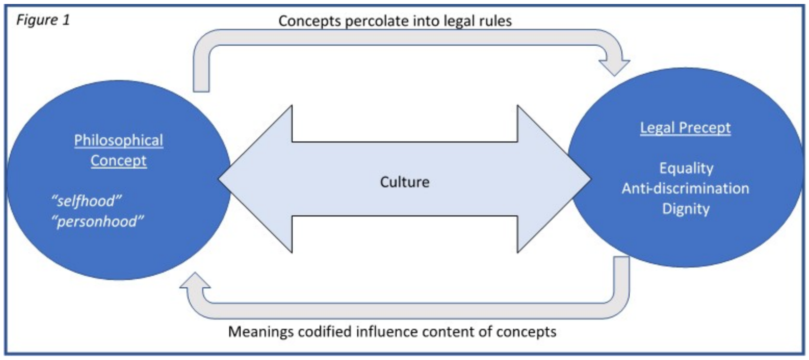
In this Part, I hope to help us overcome our occasional nearsightedness by using the insights from the previous Parts to help illustrate the ways in which legal structures exist within the context of a dynamic discursive process that translates indeterminate philosophical concepts into agreed-upon legal standards. Recognizing this translation process will help us to take note of the philosophical undercurrents underlying some of our legal precepts and encourage us to be conscious of the ways in which the sociocultural narratives surrounding such concepts can feed into our legal structures.

To begin with, consider Figure 1 below, which represents the dynamic interaction between philosophical concepts, cultural discourses, and legal rules.

[185] Indeed, it's possible to conceive of such a rule as an enactment of the principle of equal moral worth itself, considered against the background of ultimate epistemological uncertainty about the ipseity of others.

Philosophical concepts, which appear on the left, often represent objects of inquiry whose nature is essentially contested. Selfhood is one such concept: what the self is, whether it is real, and how to define its contours are all on-going philosophical quandaries.[186] Though there are no finalized answers, the notions themselves nevertheless percolate to inform our legal precepts—often our most important ones. For instance, provisions like the equal protection clause[187] and anti-discrimination regulation[188] arise, as we have seen, out of notions that deploy these concepts in particular ways. Anti-discrimination provisions might, for example, rely upon the notion that all people are "conscious beings who necessarily have intentions and purposes,"[189] and thus that the law should prohibit actions that "suppress or destroy consciousness."[190] But the translation is not just one-way. The shape of our legal rules, the ways in which those rules define and deploy the philosophical concepts, can also feed back into the meaning of the concepts: driven by the chosen application of legal rules, we might begin to conceptualize persons as, for instance, that set of selves that is capable of wielding certain rights, subtly modifying the underlying notion.[191] The mode of translation comes in the form of cultural discourse: though the same concept might possess different meanings in the philosophical and legal contexts, both meanings—likely in a somewhat diluted, imprecise form—ultimately make their way to the culture at large, where regular usage results in a sort of common denominator of meaning. The common conception of a self might, for example, be some amalgam of the idea of being a conscious, autonomous being with basic human dignity.[192] Because of this interactive relationship, the development of cultural phenomena that impact upon philosophical concepts can influence the legal approach to novel problems—like, for instance, the confrontation of seemingly-sentient automata.

Figure 1



[186] *See, e.g.*, DAINTON, *supra* note 3.
[187] U.S. Const. amend. XIV.
[188] *See supra* note 161.
[189] Réaume, *supra* note 174, at 20.
[190] *Id.*
[191] *Compare* Niman, *supra* note 150.
[192] *See* Réaume, *supra* note 174, at 20–21.

To chart how this is the case, using the introduction of automatons as a novel legal problem, examine Figure 2 below.

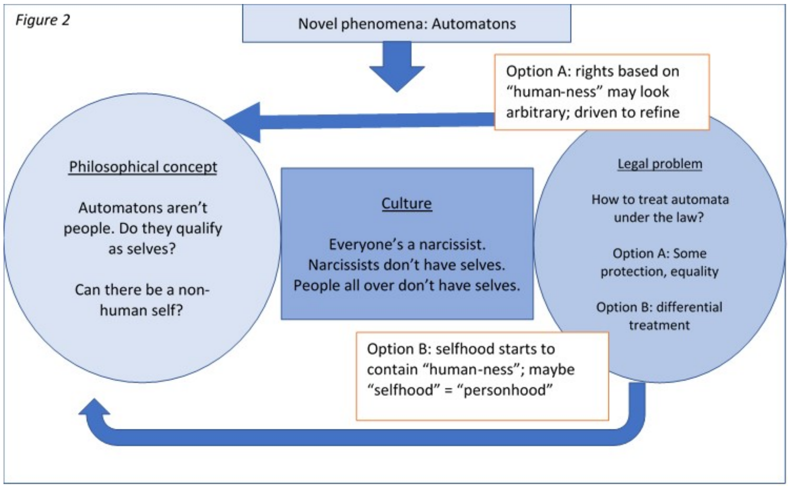


Figure 2
Novel phenomena: Automatons
Option A: rights based on "human-ness" may look arbitrary; driven to refine
Philosophical concept
Automatons aren't people. Do they qualify as selves?
Can there be a non-human self?
Culture
Everyone's a narcissist. Narcissists don't have selves. People all over don't have selves.
Legal problem
How to treat automata under the law?
Option A: Some protection, equality
Option B: differential treatment
Option B: selfhood starts to contain "human-ness"; maybe "selfhood" = "personhood"

Displayed in the middle is the existence of a cultural discourse that exacerbates skepticism about selfhood and consciousness. The bubble on the left represents the role of philosophical concepts; the one on the right represents the role of legal precepts. As the novel phenomenon of automatons enters the scene, the problem of widespread skepticism about selfhood, a problem that splits "selves" from "people", might feed Option A: the recognition that automatons have "selves"—that there can be non-human selves, that apparent consciousness is sufficient for selfhood. This kind of treatment could, in turn, impact the underlying philosophical concept by problematizing it further: if selfhood is split from personhood, what does it mean to be human? Could concepts like human dignity become no longer sensible? How far would this paradigm require recognition of moral worth to expand? It could, alternatively, drive Option B: if automatons are, in all relevant external senses, indistinguishable from humans, but the emptying out of the concept of the self provokes a turn to reliance on "human-ness," differential treatment could result. The reality generated by this legal resolution would, in turn, impact the philosophical concept: the notions of selfhood and personhood might also begin to narrow. In this way, we can see that legal structures don't exist in separation from indeterminate philosophical questions, but instead in conversation with them.

Conclusion

In this Article, we have asked a lot of imaginative questions. In Part I, we examined the ways in which confronting future Automatons might resemble confronting Bad Ex-Boyfriends, in the sense that both provoke similar psychological responses and prompt similar philosophical inquiries. In Part II, we explored the narcissism script in order to scrutinize the ways in which sociocultural forces can constrain and impact the tools available to deal with the terrors and quandaries presented by confrontation with a separate consciousness. In Part III, we introduced the thought experiment of Automatons Amongst Us, and projected three possible Scenarios for how the narcissism script might mediate our treatment of automatons in the future. Then we considered whether any of these scenarios might constitute discrimination under two legal theories. Finally, in Part IV we investigated the broader implications of the case of the automatons, asking what it might show us about the overarching process by which philosophical concepts inform and interact with legal precepts.

There is, then, between the fictional exploration of automatons in the popular imagination and the legal articulation of automaton rights designed to form actionable principles in the real world (of both the present and the not-so-distant future), both a conceptual and an epistemological distance. But this does not mean that bringing imagination to the task of exploring robot rights is fruitless; very much to the contrary—it is both necessary and useful. Necessary because the metaphors that we use to understand ourselves pervade and percolate within all of our social institutions: we bring ourselves, our existing cultural frameworks, and our narrative devices to the making and interpretation of law. This is perhaps even more pronouncedly the case in our modern, increasingly technologically-advanced societies, which present new legal challenges at a rate that outstrips the sociocultural evolution of myth—leaving us to apply the metaphors we have to the ever-brave-new-world developing around us. And useful because mining imagination can yield insight into the role that culture plays in the ongoing translation of philosophical concepts into legal precepts. In other words, when we craft laws, we draw from concepts that are culturally informed and discourses that are sociologically contingent. But it is easier to see this when we step outside our own experiences, changing just a few details of the setting and asking what difference, if any, those changes may make.