HRI ethics and type-token ambiguity: what kind of robotic identity is most responsible?

Thomas Arnold email: thomas.arnold@tufts.edu
Matthias Scheutz email: matthias.scheutz@tufts.edu
Human-Robot Interaction Laboratory
Department of Computer Science
Tufts University
200 Boston Avenue
Medford, MA 02155 USA

Abstract

This paper addresses ethical challenges posed by a robot acting as both a general type of system and a discrete, particular machine. Using the philosophical distinction between "type" and "token," we locate type-token ambiguity within a larger field of indefinite robotic identity, which can include networked systems or multiple bodies under a single control system. The paper explores three specific areas where the type-token tension might affect human–robot interaction, including how a robot demonstrates the highly personalized recounting of information, how a robot makes moral appeals and justifies its decisions, and how the possible need for replacement of a particular robot shapes its ongoing role (including how its programming could transfer to a new body platform). We also consider how a robot might regard itself as a replaceable token of a general robotic type and take extraordinary actions on that basis. For human–robot interaction robotic type-token identity is not an ontological problem that has a single solution, but a range of possible interactions that responsible design must take into account, given how people stand to gain and lose from the shifting identities social robots will present.

Introduction

In his famous consideration of the mechanical reproduction of art Walter Benjamin remarks, “Even the most perfect reproduction of a work of art is lacking in one element: its presence in time and space, its unique existence at the place where it happens to be” (Benjamin et al. 1970). Few would view a social robot primarily as a work of art, but Benjamin’s simple observation touches upon a key ethical point for the study of human–robot interaction and robotic design. Robots are not special for being manufactured or “mechanically reproduced,” of course. Unlike the Mona Lisa or Starry Night, there need be no one original, authentic robot from which all others derive (though the Apple museum, for example, testifies to how the human thirst for original relics even applies to technological devices). On the contrary, the very promise of robots to perform reliably, predictably, and effectively seems to hinge on each one of a certain kind being made to the same specifications. What makes Benjamin’s remark apposite, instead, is the crucial role of presence. Social robots do not hang on a wall but function as mobile, interactive systems, sharing time and space with particular users and interactants. It is clear that interactions with a robot can elicit great emotional investment from people, from military funerals to household gifts (Carpenter 2016; Scheutz 2011).

The opposition between robots being dependable and constant, yet socially adaptive and engaging, creates tension in how one identifies robots in the midst of interaction. When does one discrete robot of a certain model of household robots take on definitive individual characteristics, either from the perspective of its interactants or the nature of its learning design? How should its particular situation and interaction history define it as a particular machine that can only be imperfectly replaced? What particular qualities should develop for a discrete, individual robot, as opposed to all the others made according to the same design specifications?

The oddity of robotic particularity is nothing new in science fiction, of course. TV series like Battlestar Galactica and Westworld, to take but two examples, show how disorienting it can be to see robots as copies vs. unique individuals. Projecting memory, identity, and self-determination onto robots often trades on intuitions about agency and dignity. In such stories, granting personhood or moral status, especially as a victim of injustice or cruelty, seems dependent on granting particularity and uniqueness to their experience and life story. Still, these works of fantasy touch on only some of the pressing and practical questions now faced by the real-world study and design of human–robot interaction. Given the many important contexts in which social robots are either operating or being imagined to operate soon, there are foreseeable challenges to how a robot's role relates to personal history, emotional attachment, functionality, and the efficiency of replacement. And if robots are going to be morally competent (Malle and Scheutz 2014), meeting these challenges means anticipating people's reactions and vulnerabilities to oddities of robotic identity. It also means making sure autonomous robots arrive at and explain decisions that acknowledge those features.

This paper treats a subset of these emerging challenges for the purpose of developing and implementing social robots more responsibly. Using the philosophical distinction between "type" and "token," we distinguish type-token ambiguity from larger dynamics of indistinct or shifting robotic identity. After setting out the type-token challenge in contrast to distributed or networked agency, we explore three main areas of human–robot interaction (HRI) where a robot's peculiarly elusive status as a general representation (type) and a particular instance (token) takes on ethical importance: (1) personalization through interaction, especially on longer-term bases; (2) justification and self-identification by the robot, especially for moral appeals; and (3) the practice of replacement, both for high-risk or urgent situations and for longer-term relationships with a user. Each of these areas, we argue, is a likely site of tension that robotic designers must manage

for the responsible application of robots across many contexts. The purpose of addressing type-token complexity is not a final determination of what a robot is or is not, but rather determining what it ought to do with the most well-considered, wisest relationship to human needs.

The type-token distinction

The type-token distinction first appears in explicit form in the semiotics of C.S. Peirce, though in terms of intellectual history it can be viewed (as Peirce himself testifies at points) as a continuation of medieval philosophical interest in the universal vs. the particular. Peirce delineates a sign as being a type inasmuch as it is general in occurring across different instances, as the same word "the" might be seen in different sentences. A token is what Peirce calls the "*hic et nunc"* instantiation of a type: the here and now particular word "the," composed on a particular line and page, with pixels or ink, in a physical site (Peirce 1998, p. 488). The difference between the more abstract type and tokens as its "spatio-temporally located particulars" (Wetzel 2009, p. 5) enables linguists and philosophers of language to articulate formal distinctions around language, objects, and reference; for example, whether types are best represented by set theory (tokens belonging to a type as members of a class) or as a kind of rule (i.e., a rule followed by tokens) (Wetzel 2014). The type-token theory has spread from Peirce's thought toward other philosophically inflected topics. Philosophers of mind have called upon the notion of type to designate a mental state that could (or, for opponents on the other side, could not) be realized by different tokens, i.e., particular physical system states or processes.

Social robots, along with many other examples of machines, share some everyday connections to type-token judgments and classifications. If someone asks me what I'm driving I can refer to a make and model as type, but I could also refer to a

particular token of that model (e.g., the very car I took across country last summer). For robots that are interacting with and serving people in social spheres, being both a model and instantiation of the model can possess such a dual identity through the history they accrue. In addition to its unique spatio-temporal location relative to other robots of its make and model, a robot may undergo events that shape it, or mark it, distinctively. These may be physical and exterior alterations (e.g., damage or decoration) or, via social interaction, information about people or surrounding environment that the robot is counted upon to know and draw upon in its communication and decision-making.

Before moving into how those "tokenizing" processes might best be anticipated, it is important to acknowledge how recent work in HRI and related philosophical treatments of robotics has perceptively pointed out that social robots need not be viewed as isolated artifacts (Suchman 2006; Seibt 2017). Rather, robotic technology can be envisaged and evaluated as components of a larger interactive process. If the type-token distinction is meant to be a means toward ontological determinations of what a single robot is (i.e., type or token?), then the more holistic view of robots could obviate having to make them. Given how many science fiction narratives have dramatized various ambiguities about what a robot is, it is also worth putting well-worn intuitions about individual agents (particularly when projected as universal by the West) to the test of real-world scenarios and adaptive design.

Our turn toward the type-token distinction, however, has neither ontological finality nor a rigid model of agency in its sights. The ethics of human–robot interaction has quite enough rich material to sort through before resolving questions of consciousness or selfhood, especially for what a robot in the abstract is or is not. Instead, the distinction is a way to organize and anticipate interactive dynamics toward making a down-to-earth difference in robotic applications. What is crucial is not what a robot is *in esse* but its function with and impact on people,

from how robotic physicality is perceived to how robotic systems will communicate and act for society's good. A robot's interactive identity is both pragmatic and phenomenological terrain. Encounters that invite others to regard the robot as unique or separate—identifying damage on the body, movements it has learned from previous interactions, information about a user that it can bring up—may enhance or hamper the good a robot can perform. In what follows our focus is how type-token tensions can emerge in and shape real-life interaction.

The prototypical frameworks for human–robot interaction, from public articles to a great deal of scholarship, is the individual autonomous robot interacting with a person in close proximity. The social space of interaction is dialogical, often oriented toward either therapeutic or general companion-based service. The challenges of creating a positive interaction in a set context are about rapport, trust, effectiveness, pursuing interests of the human being, and others. By way of expanding that approach, the following section examines robotic identity in light of distributed systems and the dynamics of networks. From that wider vantage point the type-token ambiguity can more clearly situate real-world challenges for human– robot interaction. Not surprisingly, each of the three areas we then explore will raise more questions about robotic identity than the type-token distinction alone involves (e.g., diffuse or vague boundaries between robots and each other, or systems blurring the line between human and robot decision-making). While we attempt to keep its limited scope as a heuristic in view, the distinction still serves to organize and track an important set of practical demands.

Distributed agency, networking, and fuzzy robotic identity: situating type-token ambiguity

Before looking at where type-token challenges promise to crop up in HRI, we should pan back to appreciate that demarcating the identity of a robotic system can be challenging for many different reasons. The solo Terminator figure is an inadequate icon for how robotic platforms, not to mention automated systems overall, promise to work in the world. This is in part due to developments like machine learning's multifarious presence in finance, law, and education, not to mention the introduction of augmented reality into public space. IBM's Watson activates a world of shared expertise and perspective for real-time deliberation and decision—if Watson is operating on a robotic platform, where is the "robot's" expertise? In the case of Amazon's robot workers, a group of robots sharing information and navigating in concert does not seem accurately described as many (individual) robots but rather as one multi-part system.

No doubt robotic systems can be more useful through being diffuse and distributed, but different forms (alone and in combination) will have varied connections with their wired and unwired surroundings. Not only could one robot's learning transfer to all other robots more quickly (Scheutz 2014), but complex decisions could be made while incorporating as wide a range of information as possible. Instead of the question being whether a robot acts as a type or as a particular body, the question could be in what kind of network or "hive-mind" it operates as a contributor, as Google's recent project on robotic learning suggests (Knight 2016); moreover, we might ask how well we can manage our own actions in concert with devices whose information-sharing or processing so exceeds and defies the limits of our own (Williams et al. 2014). Thinking in a more process-oriented fashion about HRI can rightly loosen assumptions that robots represent one-to-one anthropomorphic replacements or surrogates for people, and such thinking can parse social behavior more carefully as kinds of simulation (Seibt 2017). Looking more critically at what

is genuine vs. simulated in robotic behavior (Bickhard 2017) one might deem the type-token distinction an unnecessarily isolated vantage from which to describe robotic systems: the nature of the whole simulation between human and robot may be more salient to explore than how individually distinct a machine is as opposed to being a mere shell for a general program. From a sociological vantage point, one might also underscore how social robots' individuality has been part of casting them in replacement roles, even for very delicate and intimate care work (Suchman 2006). The type-token problem might seem less relevant once one imagines robots in a range of supplementary roles, instead of being a one-on-one alternative to person-to-person care.

It is, however, precisely the embodied and multi-layered character of human–robot interaction that justifies a closer look into the type-token distinction and the pragmatic ambiguities it may yield. This distinction serves as a heuristic for anticipating how interactions might unfold, not a clean theoretical determination as one might more likely find in mathematics or philosophy of mind. Within HRI, a type-token challenge has to do more with people's attributions, responses, intentions, and plans than the robot itself as an artifact. Any critical scrutiny toward genuine vs. simulated interaction (what a robot only "pretends" to do) must still acknowledge with the empirical realities of plausible attributions that HRI scholarship continues to explore. While particular robots facing off with human beings may not capture the complications of system-wide dynamics and distributed agency, work in HRI suggests that embodied interaction can elicit a number of organizing attributions from the human psyche. Coordination and rapport on a person-to-person level has certain orienting assumptions: what we can each perceive in shared space, what kind of information is sharable, and how it might direct joint exploration of the environment. These instincts and abilities can carry over into how robots are perceived and received (Strait 2013). What kind of affect, intentions, and reasoning are inferred from robotic performance, and what

inferences are most productive and helpful for a robot to invite, will in part define what kind of interactive artifact a robot manages to be (including how it might be legally categorized). (Calo 2015). A more holistic view of interaction does not preclude addressing uncertainties of robotic deployments to which we have been shown susceptible.

Again, in the three areas of HRI we flesh out below, the type-token ambiguity is important for how it affects future interaction. What a social robot of one type (body, form, operating system, control architecture, etc.) or another ought to be called, in and of itself, is left for other philosophical treatments. In the interplay between real functions and generated attributions, robots' general and individual features ought to assist, not interfere, with their fullest service within their application domain. Like many facets of human–robot interaction, the areas below serve both as conditions under which a problem arises (in this case type-token ambiguities) and a practical feature of robots within which the problem might need to be managed.

Personalization, memory, and shared history

Across various contexts of human–robot interaction, the type-token distinction can track how a robot establishes a unique presence (through physical qualities, abilities, and memory) while also representing a template or design that other robots could embody. As robots become more personalized to their users, and perhaps also other people around them, this can guide the imagination for what kind of relationship can develop (Sung et al. 2009). In contexts where a robot is communicating and serving people, there will be particular adaptations a robot could need to acquire if its social skills are to sustain practices of caretaking, teaching, and guidance. Responsive dialogue and episodic memory of past exchanges seems to be a prerequisite of companion robots. For if social robots

remain “types”, i.e., if all instances remain the same irrespective of the contexts and individual interactions (as is, for example, the case with vacuum cleaning robots), they will likely disappoint their users who are seeking to connect with them on a personal level (Breazeal 2002). Assistive robots, which carry out tasks that people deeply appreciate, might appear indifferent if memories do not alter their behavior, causing frustration if not emotional harm due to their perceived ignorance (which, in the longer term, will likely hurt their task performance). For when a particular robot has multiple interactions with a person that draw on modalities of intimacy (natural language, facial expressions, gestures, bodily positioning, responsive movement and orientation, cooperative physical action, shared space), these interactions may be construed as carrying longer-term relational aspects.

If a robot can adapt to personal qualities, needs, and habits, it may accomplish more for a person in a number of ways. Robots that have been told personal histories, or have shared a moment the user deems remarkable, could serve as a memory aid, as well as a basis for future suggestions and appeals. The recent movie *Marjorie Prime* explores the idea of an AI system helping a user with dementia (though the system is holographic projection, which lacks the solid physical presence of a robot). There may be subtle movements or behaviors that a robot learns that the person prefers (e.g., around how to help them out of the bath), mastery that is then expected from this particular robot because of a shared learning process together. The idea that such knowledge could be transferred to many other robots without much delay, as discussed earlier with respect to networking, may be one uncanny or creepy part of a robot’s identity. And the type-token confusability could crop up around struggles with attachment and bonding, where having a functional robot is not as good as having *this* robot (with distinguishing marks) that went through special times with the user. Nor would social actions of a robot need to be extraordinary at all in order to have

thoroughgoing effects on its relationship with people. Remembering even simple events or facts, especially presented as secrets, can draw out a host of interpersonal judgments and attributions that might mistakenly attach to the robot (Kahn et al. 2015).

Personalization brings out another interesting type-token tension when it comes to situations of ethical judgment on the part of the robot. If an interactant breaks a rule or violates a norm of some sort (refusing doctor's orders, damaging teaching materials, lying about an event only the robot saw), the robot may face competing priorities in its operation. On the one hand, there are basic rules, prohibitions, and standards that a robot would need to uphold to help provide a stable, reasonably safe working environment for those around it. At the same time, a socially engaging robot must be more than a rigid taskmaster. Not perceiving mitigating circumstances (illness, grief, ongoing work to change, etc.) risks repelling the very people the robot is designed to help, undermining the effectiveness of its moral appeals. Especially for caregiving or coaching roles, there may also be an expectation of solidarity or shared purpose, such that calling out failures does not serve an overall project of improvement and growth (Demiris 2009). On the other hand, simply ceding to what a person decides, no matter how bad the action, undercuts the robot's ability to serve the community beyond them and ultimately its standing as a normative, if not moral agent.[1] Failure to point out normative transgression and thus to express blame appropriately will also alter the human perception of the robot's capability and autonomy, with its own set of ensuing effects (Briggs and Scheutz 2015). And it will consequently raise the question

[1] Note that we specifically refrain from engaging in the ongoing philosophical debate as to whether robots *can* or *could* be truly or fully moral agents; all we indicate here is an interactant's construal of the robot as norm-following or moral.

whether the robot represents a systematic decision-making type, or a token that has crucially customized qualities and skills.

Type-token dynamics are not dependent on an overly anthropomorphized form. People who use Siri understand a token function in that Siri will answer the question they ask then and there on their device, yet they recognize Siri as a type who will answer a friend's question in a similar way. With sufficient customization, this could become more problematic. The movie *Her* imagines an operating system that reaches much greater levels of intimacy, replete with convincing shows of affection and desire. In this case the user is dismayed to find out he was mistaken about the system's identity: its thousands of other user relationships showed that the system was not confined as an isolated entity, but more like an overarching type that instantiated for each user. The case of social robots in real-life contexts may not be as stark as that storyline, but possible forms of type-token ambiguity could put strains on people's various social impulses. Familiarity with a person's behavior and actions, including their narrated memories, could be as distancing as it is engaging, depending on whether a person feels more exposed and vulnerable than acknowledged and understood.

Type-token forms of personalization could also take advantage of implicit interaction, identified and explored by Ju (2012) as a set of underappreciated dimensions of robotic embodiment. The power of touch, gesture, expression, bodily orientation, and shared activity engages more of our social capacities to attribute identity (Yohanan and MacLean 2012; Farmer and Tsakiris 2013). When one considers how touch, temperature, texture, and proxemics can shape human–robot interaction—especially where intimacy and power shape eldercare, education, medicine, and public space—personalization and standardization could stoke or defuse feelings of loss, fear, unfairness, and safety.

When one adds back in considerations of networking and institutional interfaces (where other people not directly interacting with a robot may still have access or control to the robot, e.g., in a medical facility), there may be even more awkward and unsettling moments of personal information being shared and utilized beyond the direct human–robot interaction where it emerged. When one further remembers the robot's own physical vulnerability, its susceptibility to damage, surface marking, or other alterations, it is clear that a particular robot may take on a highly personalized role for a person with whom it interacts, from physical appearance to higher-order appeals to self-improvement. A child who wants one certain robot as tutor again because of a great lesson it once taught; a resident at an eldercare facility who counts on one robot to recount fading, but private memories but not share them; a therapy robot that knows exactly where certain injuries of a patient are, but can also impart that knowledge to any other robot in the facility—these are complex scenarios where robotic identity is not solved like a technical problem but managed and refigured through the course of relationships (Caine et al. 2012).

Justification and self-presentation: moral appeals and system transparency

The personalized interactions that successful social robots might provide entail more than retained details here and there from a user. They can also extend to judgments and appeals about what should be done. When it comes to observing rules, noting personal frailties, judging reasonable measures, a robot would ideally not fall into ruts of unyielding enforcement or empty complicity. The problem, however, extends beyond just whether or not a robot chooses to enforce a stated rule, as well as what kind of information counts as mitigating circumstances in the personal experience of the robot's interactants. The type-token distinction is also a problem about justification. The robot's own means of decision-making, either as a

shared appeal to norms or a computational process a person may want to know more about, may emerge as a bone of contention in a longer human-robot deliberation. Two intertwined questions arise on the basis of thinking through how a robot justifies its action: (1) How does the robot decide? (2) How should it represent its decision-making process?

The type-token distinction adds its own twist to these questions. Are the reasons this robot is using based on distinctive interactions as an individual robot, or is its ethical judgment a process indistinguishable from any robot of its type? Moreover, when is it mandatory for the robot to represent its own computational approach to an ethically challenging situation, and when should it deflect curiosity about its processing in order to reinforce an appeal strictly to explicit norms or values (Arkin et al. 2012)?

It is beyond the scope of this paper to explore the full range of issues around justification and machine ethics. Ethical theory and computational architecture have seen some important developments with respect to addressing moral challenges, for example, hybrid approaches to difficult decisions in morally charged scenarios. Transparency of autonomous systems is unquestionably a practical must for embodying moral norms in the world, since such norms almost always hinge on socially appropriate communication and coordination with others. Yet, justification does face a particularly knotty problem when viewed through the type-token dichotomy. For when an autonomous social robot reaches decisions with and among people, it will face serious complexities of self-representation and moral agency in the course of giving and receiving reasons. It may serve as an "explicit moral agent,", yet need discretion (Moor 2009). Consider a question a resident might ask of it in an eldercare facility: "Why should I go outside and take a walk?" An eldercare robot's answer can have several levels of specificity and argument. It could emphasize general reasons like the fact that walking is good for circulation, but it might also enlist personalized reasons like "You didn't exercise

yesterday." But let us say that recommendation gets more complex, for example, that there is threatening weather outside. A person might want to know what exactly this robot is considering, not what every model is programmed to say. The person, in other words, may want to know how particular its reasoning is, i.e., "Why do YOU [this robot in front of me] think I should walk?" Again, there is a general problem about transparency and moral reasoning that work in HRI must work to resolve in practical terms; that is to say, when does a robot represent the external reasons for an action as opposed to its own computational means by which it arrives at those reasons? What type-token ambiguity helps identify more exactly, though, is how familiarity with a robot might deepen users' confusion or consternation about what the robot's stated reasons fundamentally reflect: a systematic priority/value or vs. a personalized, distinctive output. If the former, then the user may have to pursue whether enough contextual information is supplementing the robot's decision ("Have you looked outside?"), including personalized information the resident might expect the robot to know ("I said yesterday I can't get too much sun"). If the latter, then the resident might still question what use is being made of that (i.e., is the robot trustworthy to "do the right thing" regardless of their impulses).

Personalized justification, especially in contexts of professional authority where the withholding of some information is assumed or suspected (e.g., medical cases), the robot's particularized answers may only exacerbate a sense of processing "behind the curtain," a justification behind the stated reason.

Are there symptoms the robot should take into account but whose declaration to the resident would unduly upset them? As people become savvier with computers, coding, and AI, pivoting from a shared moral question ("What should I do?") into a system diagnostic ("Why is your programming saying that is the right thing to do?") seems quite plausible as a common scenario to anticipate. For that reason, computational approaches for moral competence will have to do more than

incorporate a simple set of natural language justifications. There must be some accounting for the layered identity of the social robot. One might point out that people serving in delicate social roles (especially as caregivers) can have divided identities. One can act as an individual, a professional, or a representative of an institutional policy, with many possible conflicts between roles. There are ways that trust and knowledge are built in recognition of wearing multiple hats—a doctor can communicate a level of personal sincerity to their professional judgment by saying, "If it were my child, I would go see this specialist." But a robot has no common interests or vulnerabilities (like children, or personal suffering) to represent in good faith, at least in the real world (science fiction stipulates them and asks us to imagine accordingly, with suspended disbelief). What the robot could have instead are the interests it is designed to serve through computational and physical means. Upon being questioned it has a more definite kind of accounting to provide: what it is designed to conclude is best, the reasoning it uses to reach that conclusion, and the information and assumptions by which it does so.

There would appear to be a tradeoff between a robot sticking to explicit reasons without further justification and divulging systematic processes that the user might misinterpret or mistrust. That is to say, a robot's moral competence in performance might be constrained by the trust its justifications can support and sustain. Computational accommodations of these type-token shifts will have to adjust and acknowledge the different ways that justifications can work. A kind of biographical representation of the robot's interactions might make sense, as a means of both accountability and transparency for what the robot can and cannot judge dependably. At the same time, the robot's justification for actions should be geared, as much as possible, to meet the service purposes of those deploying it, not to overwhelm end users with internal detail. The type-token complexity is not resolved or solved once and for all, but is to be incorporated humanely into design and deployment.

The goal would be an intuitive, reliable type-token awareness on the part of all stakeholders: how an individual robot's interactions relate to that of others that could serve in its stead, the "village" of moral standards that stands behind it and must keep it accountable.

The practice of replacement: risks and opportunities

Ordinary associations with a computer or tool to be replaced have to do with either disrepair or obsolescence relative to a more advanced product. Personalization of various kinds, for reasons we have seen, may elicit more resistance than normal from a user toward a particular robot being swapped out, even for one of the same model. Emotionally vulnerable interactants, especially those with difficulties tied to loss or sudden change, are prime people to consider when implementing robots for important social roles.

Replacement carries even more ethical challenges and opportunities when one takes the perspective of the autonomous system reaching decisions within critical, high- stakes situations. In contexts where a robot's work does not depend primarily on direct social interaction, its actions may affect large numbers of people, who may have little time or perspective on what the robot should do. Recently some attention has been paid to how robots might go "above and beyond" prescribed tasks (Bringsjord 2015; Scheutz and Arnold 2016). Not surprisingly, artistic renderings of robotics can cast this in a heroic or sacrificial frame, for example, the recent movie *Big Hero 6*. There are, however, good reasons to anticipate scenarios of imminent danger where robots might play a critical, unique role due to their abilities and purpose (Lin 2016).

A road repair robot, for example, might be expected to extend beyond patching a pothole if a child ran in front of oncoming traffic. In situations of dire human need,

where a robot could easily be destroyed or permanently incapacitated, its status as token of a robotic type might well shape what makes sense for it to do. With certain dramatic, risky actions, the fact that a robot could have an easily reproduced identity would suggest the action is worth undertaking (this is perhaps part of why science fiction so often involves robots on impossible missions faced with decisions of sacrifice). If there is no likely replacement, however, then the single robot may be the only agent available to fulfill subsequent work. Unlike the road repair robot, one might think of a rescue robot that is working in an especially treacherous area of collapsing structures. It is not that the robot has interests and desires for its own sake; it is that risking permanent damage might prevent its ongoing service for many others needing rescue. However such utility is approached computationally, it is clear that a robot's peculiar conditions of identity affords more than a feeling of uncanniness, unsettledness, or disappointment – lives may be at stake depending on how the robot's computational architecture can represent that perspective properly and ethically. Such "above and beyond" actions in critical moments are, nevertheless, still ethically tethered to more ordinary contexts. What if a personalized robot has built an unusually strong, productive relationship with a person? Would that change the conditions under which that robot might sacrifice itself? One might also think about the robot's particular body: the distinguishing marks or material quality that an interactant would come to know intimately (as with tools) through time. How would that affect seeing it demolished or replaced, both in observing it act and coping with the loss going forward?

While its replacement could be a matter of just transferring code and resuming life with another copy, it could also run against the grain of the original's robot idiosyncrasies, from physical marks or defects to particular qualities its material has from wear. Those worn-in features are evidence for the physical reality that this object, not a replacement, was present for this or that experience with a person.

And it means that replacement, whether planned or unplanned, has to account for the kinds of dissociation various users would have to carry out.

On a material and cultural level, it is difficult to see human–robot interaction being exempt from the same affects, practices, and rituals by which keepsakes and relics try to encapsulate and bear witness to past physical reality (though ethnographic evidence on Aibo and Roomba robots suggests that this is not so, e.g., see Friedman et al. 2003 or; Sung et al. 2007). To be sure, innumerable computers and smartphones lie unmourned in landfills—dissociation and disposal are not always difficult for us when it comes to technological products. But given the intimacy toward which some human–robot interaction is being designed, and the contexts of vulnerability and real need that ostensibly call for robotic help, it does not make sense for roboticists to dissociate from the effects of loss and interruption that one robot's replacement could elicit. The very means of connection sought and promised as functional achievements make a throwaway social robot harder to imagine.

Directions for responsible design

As discussed from the beginning of the paper, the type-token distinction does not demand ontological resolution for its own sake. It is a practical means by which real harm and help to those interacting with robots can be anticipated and accounted for in the process of designing and implementing robots. Having outlined three areas where type-token ambiguities have ethical stakes designers and users must face, some preliminary suggestions for how to do so are in order.

The problems connected to a robot's status as token vary with its intended role and the responses of interactants. It is important to delineate and evaluate different kinds of attributions made toward social robots in terms of the intended purpose for

the robot's work. Patiency, for example, seems an especially delicate attribution for a robot to elicit (Bryson 2012). Damages to a robot's body, especially in the course of past interactions, might draw out more sympathy or feelings of intimacy from users and interactants. If that damage plays no distinguishing role in how the robot operates, especially if other robots of its type share the knowledge of those past interactions, then people do not have a true picture of what the robot and its performance represents. The resulting loss that is felt when that particular robot no longer works (or is replaced by one identical but for the damage) is one that robotic design could and possibly should try to pre-empt (examples of such cases of experienced loss, where owners wanted their damaged robot back instead of a functionally identical copy, are some owners of Roomba vacuum cleaners and former soldiers in Iraq working with teleoperated Packbots for bomb defusing, e.g., see Scheutz 2011).

There may be dissociative language and references by which a robot reminds users of their dual status. The use of first-person language to describe damage, perhaps, may not be a good default. No approach will completely avoid attachment-yielding attributions, of course; we regard old tools with memories of past work and poignant affinity for how well they have been "broken in" and suit us. Moreover, digital assistants like Siri seem to use the first-person without too much confusion. On the other hand, Siri is disembodied and counted upon to be a consistent type for anyone using the operating system—one of Siri's features, in fact, are stock answers that one can dependably hear if one asks the very same question their friend did. For the shared space and history that robots might be expected to navigate, this may well not look the same. For those in robotics and/or AI working in natural language processing and generation, due vigilance and care should be taken for self- reference, lest manipulated sympathies and outsized expectations lead to a disappointing and dysfunctional interaction.

The role of memory is another functional issue where token ambiguity deserves scrutiny and discerning foresight. Remembering an interactant's words and actions can serve many good ends, from gleaning medically relevant information to providing accompaniment and acknowledgment what has happened in a person's life. As discussed with reference to justification, it also may provide valid means of persuasion and understanding with respect to a person's interests, whether physical, emotionally, or social, as well as those of the community. As with patiency, the key is to balance the spatio-temporal reality of the single robot—it was really the one which was there, it really did receive what the person said and did—with the technological reality of how that memory is used, shared, and situated with respect to the robot's overall purpose. Emphasizing the robot as a channel, not just a self- contained unit, may be a principle for above-board interactions and healthier psychological distance for the people whom it encounters. In this way, the benefits of personalization and effective justification do not impede the realities of distributed agency and ready replacement.

Ethical dimensions of human–robot interaction, therefore, underscore the importance of tempering any projections of patiency onto robots (Bryson 2012). While there may be therapeutic ends for seeing a robot as an imaginary sufferer (e.g., children working on cause- effect of certain behaviors), overall humanoid robots can invite many more attributions than their system and function justify. Increasingly perceiving the robot token as diverging from or supplementing its type—their particular body and history over the overall function and purpose any of their type can serve—may produce ultimately harmful illusions of affection or suffering. What token identity should accomplish is memory support, and design should find means to preserve that support across individual robots. The physical interaction with a robot will unavoidably create some form of attachment, recognizing that this robot is the one that was here with me when something happened. But design must ensure that any individualization works toward the

larger purpose for which the robot is employed, especially with approaches of care (Van Wynsberghe 2013). That means preparing the way for a person to replace the robot, if not a robot of the exact same kind. It will likely mean, to the degree it meets users' needs, transferring important information and history to any successor. Funerals and medal-ceremonies for defunct robots testify to the symbolic character of robot replacement, and the design of social robots will have to anticipate such measures of loss, perhaps warning those interacting with the robot that their emotional reaction may be much more powerful than they imagine. These needs will be particularly acute for those in longer-term, social interactions, as well as those who are struggling with the cognitive and emotional aspects of attachment, memory, and relationships (e.g., dementia, trauma, etc.).

Though HRI research can justifiably concentrate on how humans react to various robotic designs, the peculiarity of robot tokens should also feature in a robot's own internal assessment for how to act in a given situation. Integrating self-identification into utility assessment should inform decisions about actions, especially high-risk and urgent ones, where the robot's status relative to its type (whether that is a general function a person or robot could fill, a function many robots could fill, or an action only others of its specific model could execute) bears heavily on what the best action might be. Some of these decisions in real-time will be too complex to resolve with adequate certainty, but neglecting the conditions of a robot's replacement could open up many more reckless, pointless, and ineffective measures on the part of robots not equipped to account for its value.

Throughout history technological change has always brought with it human beings adapting their lives to new tools, to both good and ill effect. As with the digital assistant, it may be that while some of people's behaviors with a robot might become antisocial and inhumane (e.g., verbal abuse), the oddities of robotic identity gradually weave their way into more and more familiar activities. The decided advantages of networked, distributed robotic performance may motivate

more adaptable reactions than we imagine as of yet. Interactions with robots may produce a fluid, adhoc awareness of all the ways a particular robot might be importantly distinct or reliably typical.

Even so, designers cannot shirk the obligation to keep real risks in view as robotic systems enter into social spheres. Vulnerable populations of users, especially those for whom a robot is hard to conceive of except as a separate agent, may experience confusion and manipulation if the illusion of independence and a special relationship to the user is too strong (Biegler 2016). The question is, for policy and design alike, what ultimately justifies creating that kind illusion for those users, and what can truly serve as "socially assistive robotics" by design (Tapus et al. 2007).

Conclusion

The looming threat of social robots in society often appears as the replacement of human workers, one-to-one, in a variety of occupations. Automation of certain labors across industries has provided ample reason not to dismiss that picture as mere anxiety (just look at the term "computer" itself). Still, a thorough analysis of human–robot interaction and its prospects must critically expose where replacement or duplication is not a constructive model for enhancing human action and abilities. There are asymmetries and incongruities to robotic action, including the computational means by which it decides to act, that signify more than the binary determination of a robot being able or unable to fulfill the role of a person in a certain context. In this paper, we have explored how the complexity of a robot as a token of a type can surface key conflicts around how robotic roles should emerge: the purpose of interaction, whose interests are being served, how longer-term relationships might develop, and what critical actions robots alone (and in some cases a single robot alone) should attempt.

The type-token distinction is not alien to our own ethical training and deliberation as people. Kant's emphasis on universalizable maxims goes hand-in-hand with proposing what any rational agent in a situation would do: the rational agent instantiates a universal "type" without its particular history licensing any distinctive reason, interest, or know-how to justify an action.

The concrete ways in which a robot will have to absorb information from interactions, adapt its actions in response, and plan the best action in real-time is not a technical problem with a tidy solution. Given the full-range of personal instincts a social robot might elicit from the people who encounter it—through its language, gesture, touch, and movement—it is prudent to make sure a robot does not assume an identity that its abilities and systemic structure do not warrant (Li et al. 2016). We have thus specifically stressed how robots test the societal question of how individual actors best represent general values in their communities. Robots do so not only because, as machines that might be networked or transfer programming, they unsettle our usual intuitions about situated individuals facing ethically charged predicaments. They also introduce considerable ambiguity as to what relationship they should have with people, whose expectations and habits may lead to delusion, disappointment, and adverse dependence.

However adaptive and technologically enmeshed human beings prove to be with robots, the empirical results of HRI study testify to the robustness of embodied interactions with social robots. Weighing the advances and opportunities for robotic assistance against the social risks of dehumanizing social effects will be an ongoing, on-the-ground task for those who closely attend to the "holistic interaction" that humans and robots effect (Young et al. 2011). Moreover, the robots' dual presentation as agent and artifact will not change with technological familiarity alone; when it comes to justification, facility with robots may produce even more complicated social conditions under which a robot should reveal its computational underpinnings. Responsible design and policy is better served

anticipating and preparing for how such problems may develop than hoping they will pass us by.

References

Arkin, R. C., Ulam, P., & Wagner, A. R. (2012). Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. *Proceedings of the IEEE, 100*(3), 571–589.

Benjamin, W., Arendt, H., & Zohn, H. (1970). Illuminations; edited and with an introduction by Hannah Arendt. Translated by Harry Zohn (1st ed.). London: Cape.

Bickhard, M. H. (2017). Robot sociality: Genuine or Simulation? In R. Hakli & J. Seibt (Eds.), Sociality and normativity for robots (pp. 41–66). Cham: Springer International Publishing.

Biegler, P. (2016). The real costs of making friends with robots. The Age. Retrieved 12, December 2016, from http://www.theage.com.au/technology/technology-news/the-real-costs-of-making-friends-with-robots-20161027-gscgbe.html.

Breazeal, C. L. (2002). *Designing sociable robots*. Cambridge: MIT Press.

Briggs, G., & Scheutz, M. (2015). Sorry, I can't do that: Developing mechanisms to appropriately reject directives in human-robot interactions. In 2015 AAAI *fall symposium series*.

Bryson, J. J. (2012). Patiency is not a virtue:suggestions for co-constructing an ethical framework including intelligent artefacts. In: D. J. Gunkel, J. J. Bryson, &

S. Torrance (Eds.), The machine question (pp. 73–77). Birmingham: Society for the Study of Artificial Intelligence and the Simulation of Behaviour.

Caine, K., Sabanovic, S., & Carter, M. (2012). The effect of monitoring by cameras and robots on the privacy enhancing behaviors of older adults. In *7th ACM/IEEE international conference on human–robot interaction (HRI)* (pp. 343–350).

Calo, R. (2015). Robotics and the lessons of Cyberlaw. *California Law Review, 103*, 2008–2014.

Carpenter, J. (2016). *Culture and human-robot interaction in militarized spaces: A war story*. Abingdon: Routledge.

Demiris, Y. (2009). Knowing when to assist: Developmental issues in lifelong assistive robotics. In *2009 Annual international conference of the IEEE engineering in medicine and biology society* (pp. 3357–3360). IEEE.

Draper, H., & Sorell, T. Ethical values and social care robots for older people: An international qualitative study. *Ethics and Information Technology, 19*, 49– 68.

Farmer, H., & Tsakiris, M. (2013). Touching hands: A neurocognitive review of intersubjective touch. In Z. Radman (Ed.). *The hand, an organ of the mind: What the manual tells the mental* (p. 103). Cambridge: MIT Press

Friedman, B., Kahn, P. H. Jr., & Hagman, J. (2003). Hardware companions? What online AIBO discussion forums reveal about the human-robotic relationship. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 273–280). New York: ACM.

Kahn, P. H. Jr., Kanda, T., Ishiguro, H., Gill, B. T., Shen, S., Gary, H. E., & Ruckert, J. H. (2015). Will people keep the secret of a humanoid robot? Psychological intimacy in HRI. In *Proceedings of the 10th annual ACM/IEEE*

international conference on human-robot interaction (pp. 173–180). New York: ACM.

Knight, W. (2016). Google builds a robotic hive-mind kindergarten. *MIT Technology Review*. https://www.technologyreview.com/s/602529/google-builds-a-robotic-hive-mind-kindergarten/.

Li, J., Ju, W., & Reeves, B. (2016). Touching a mechanical body: Tactile contact with intimate parts of a humanoid robot is physiologically arousing. *Journal of Human-Robot Interaction*. https://doi.org/10.5898/JHRI.6.3.Li

Lin, P. (2016). *We're building superhuman robots. Will they be heroes, or villains?* Washington Post. Retrieved 1 December 2016, from https://www.washingtonpost.com/news/in-theory/wp/2015/11/02/were-building-superhuman-robots-will-they-be-heroes-or-villains/?utm_term=.a58657bad760.

Malle, B. F., & Scheutz, M. (2014). Moral competence in social robots. In *2014 IEEE international symposium on ethics in science, technology and engineering* (pp. 1–6). IEEE.

Moor, J. (2009). Four Kinds of Ethical Robots.*"Philosophy Now, 72*, 12–14.

Neff, G., & Nagy, P. (2016). Automation, algorithms, and politics— talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication, 10*, 17.

Pagallo, U. (2013). *The laws of robots: Crimes, contracts, and torts* (Vol. 10). Berlin: Springer Science & Business Media.

Peirce, C. S. (1998). *The essential Peirce: Selected philosophical writings* (Vol. 2). Bloomington: Indiana University Press.

Scheutz, M. (2011). The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics* (p. 205). Cambridge: MIT Press.

Scheutz, M. (2014). 'Teach One, Teach All'—the explosive combination of instructible robots connected via cyber systems. In *Proceedings of the 4th annual international conference on cyber technology in automation, control and intelligent systems*, pp. 43–48.

Scheutz, M., & Arnold, T. (2016). Feats without heroes: Norms, means, and ideal robotic action. *Frontiers in Robotics and AI, 3*, 32.

Seibt, J. (2017). Towards an ontology of simulated social interaction: Varieties of the "As If" for robots and humans. In R. Hakli & J. Seibt (Eds.), *Sociality and normativity for robots* (pp. 11–39). Basel: Springer International Publishing.

Strait, M., Briggs, G., & Scheutz, M. (2013). Some correlates of agency ascription and emotional value and their effects on decision-making. In *Affective computing and intelligent interaction (ACII), 2013 humaine association conference on* (pp. 505–510). IEEE.

Suchman, L. (2006). Reconfiguring human-robot relations. In *ROMAN 2006-The 15th IEEE international symposium on robot and human interactive communication* (pp. 652–654). IEEE.

Sung, J., Christensen, H. I., & Grinter, R. E. (2009). Robots in the wild: Understanding long-term use. In 4*th ACM/IEEE international conference on human-robot interaction (HRI)*, 2009 (pp. 45–52). IEEE.

Sung, J. Y., Guo, L., Grinter, R. E., & Christensen, H. I. (2007). My Roomba is Rambo? intimate home appliances. In *International conference on ubiquituous computing*.

Tapus, A., Mataric, M. J., & Scassellati, B. (2007). Socially assistive robotics. *IEEE Robotics and Automation Magazine, 14*(1), 35.

Van Wynsberghe, A. (2013). Designing robots for care: Care centered value-sensitive design. *Science and Engineering Ethics, 19*(2), 407–433.

Wetzel, L. (2009). *Types and tokens*. Cambridge: MIT Press.

Wetzel, L. (2014). Types and tokens, The Stanford Encyclopedia of Philosophy (Spring 2014 Edition), E. N. Zalta (Ed.), https://plato.stanford.edu/archives/spr2014/entries/types-tokens/.

Williams, T., Briggs, P., Pelz, N., & Scheutz, M. (2014). Is robot telepathy acceptable? Investigating effects of nonverbal robot-robot communication on human-robot interaction. In *Robot and human interactive communication, 2014 RO-MAN: The 23rd IEEE international symposium on* (pp. 886–891). IEEE.

Yohanan, S., & MacLean, K. E. (2012). The role of affective touch in human-robot interaction: Human intent and expectations in touching the haptic creature. *International Journal of Social Robotics, 4*(2), 163–180.

Young, J. E., Sung, J., Voida, A., Sharlin, E., Igarashi, T., Christensen, H. I., & Grinter, R. E. (2011). Evaluating human-robot interaction. *International Journal of Social Robotics, 3*(1), 53–67.