Giornale di filosofia, vol. 2/2021 • ISSN 1827-5834 • DOI 10.7413/1827-5834019

Zachary Daus[*]

Designing Mutually Vulnerable Human-Robot Interaction: Challenges and Possibilities

Introduction

Social robots are becoming increasingly prevalent in many industrialized nations. During the ongoing Coronavirus pandemic, a number of governments either began or accelerated programs to distribute companion robots to individuals suspected of suffering from loneliness, such as the elderly[1]. As companion and other social robots are utilized with greater frequency, a number of ethical concerns have been raised, the majority of which can be broadly classified into two distinct – yet often interrelated – concerns. The first is that companion robots deceive their users in ways that are morally dubious[2]; the second is that they result in a significantly lower quality of interaction[3].

This essay addresses the latter concern by invoking the concept of mutual vulnerability. I argue that mutual vulnerability is a social phenomenon that promotes both interpersonal trust as well as a form of autonomy that I, borrowing from feminist theorists such as Catriona Mackenzie and Natalie Stoljar[4], refer to as relational autonomy. Although empirical research has shown that the mere expression of vulnerability by a robot

[*] Master Student in Philosophy. University of Vienna

[1] B. Engelhart, *What Robots Can–and Can't–Do for the Old and Lonely*, in "The New Yorker", May 24, 2021, https://www.newyorker.com/magazine/2021/05/31/what-robots-can-and-cant-do-for-the-old-and-lonely.

[2] M. Coeckelbergh, *How to Describe and Evaluate "Deception" Phenomena: Recasting the Metaphysics, Ethics, and Politics of ICTs in Terms of Magic and Performance and Taking a Relational and Narrative Turn*, in "Ethics and Information Technology", XX, 1, 2018, pp. 71-85; A. Sharkey, N. Sharkey, *We Need to Talk About Deception in Social Robotics!*, in "Ethics and Information Technology", 2020, https://doi.org/10.1007/s10676-020-09573-9.

[3] R. Sparrow, L. Sparrow, *In the Hands of Machines? The Future of Aged Care*, in "Minds and Machines", XVI, 2, 2006, pp. 141-161; S. Turkle, *Alone Together: Why We Expect More From Technology and Less From Each Other*, Basic Books, New York 2011.

[4] C. Mackenzie, N. Stoljar (a cura di), *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self*, Oxford University Press, Oxford 2000.

promotes interpersonal trust, I nonetheless argue that such expression is not sufficient for the promotion of relational autonomy, which I claim is a significant characteristic of quality interaction. In order for relational autonomy to emerge within human-robot interaction, a robot must be able to both perceive the vulnerability of the human with which it is interacting and attune its behavior in accordance with this perception. Although the complexity of this procedure poses considerable challenges to designing human-robot interaction capable of promoting relational autonomy, I nonetheless argue that the interrelated concepts of mutual vulnerability and relational autonomy can help guide policy makers in implementing a more humane utilization of companion robots.

The first section of this essay consists of an explanation of the concept of mutual vulnerability. The second section analyzes how mutual vulnerability encourages trust between humans and how previous literature, with the exception of Thomas Hobbes, has largely overlooked the role of mutual vulnerability. The third section analyzes how mutual vulnerability encourages a form of autonomy that I refer to as relational autonomy, arguing that a similar conception of autonomy is already present in the political thought of Hannah Arendt. Finally, I discuss the challenges of designing robots capable of promoting relational autonomy and suggest ways in which the concepts of mutual vulnerability and relational autonomy can nonetheless be incorporated into the humane implementation of social robots like companion robots.

Mutual Vulnerability

Broadly defined, mutual vulnerability is any situation that consists of two or more individuals vulnerable to the same risk. Consider the example of the crew of a small sailing vessel at risk of sinking in a storm. In this case, the crew is mutually vulnerable to the risk of their vessel's sinking, due to the fact that all are corporeally vulnerable to the risk of drowning. When individuals are individually vulnerable, on the other hand, they are vulnerable to separate risks. Consider the example of an individual in war and an individual in peace. Although both are corporeally vulnerable, one is vulnerable to risks like enemy fire while the other is vulnerable to risks like traffic accidents.

Mutual vulnerability manifests itself in the form of corporeal vulnerability seemingly rarely, for the most part isolated to extreme situations of collective endangerment, like that of the aforementioned sailing vessel. One manifestation of mutual vulnerability that seems to appear more frequently than that of corporeal vulnerability is what could be called professional vulnerability. Often colleagues must collaborate on

joint projects such as joint publications or joint research. When two or more colleagues commit to such a project, it is possible that they become professionally vulnerable to the risk of the project's failure vis-à-vis the potential failure's negative effects on their careers. It also seems that our mutual vulnerability to a shared risk can be simultaneously conditioned by a number of distinct vulnerabilities. For example, if the colleagues of a joint project are also friends, they might also be interpersonally vulnerable to the risk of the project's failure vis-à-vis its potentially negative effects on their friendship.

Lastly, it is worth mentioning that the vulnerabilities of mutually vulnerable individuals, while necessarily having the same *risk*, do not always have to be the same *vulnerability*, in the sense that the vulnerability of the crew of the aforementioned sailing vessel is universally corporeal. A doctor and a patient, for instance, can be mutually vulnerable to the risk of a treatment's failure but in different senses: the doctor in a professional – and legal, if medical malpractice was committed – sense; the patient in a corporeal sense. As long as the different vulnerabilities share the same risk, it is sufficient for some degree of mutual vulnerability to emerge.

Mutual Vulnerability and Trust

In this section, I provide an explanation of how mutual vulnerability helps engender trust as well as an overview of how past literature has largely overlooked the phenomenon of mutual vulnerability. First, I begin with a summary of the notions of trust commonly used by engineers when designing trustworthy robots, the majority of which exclude any account of vulnerability. Then, I provide a brief overview of recent vulnerability-centered conceptions of trust offered by social theorists. Despite their focus on some notion of vulnerability, I argue that all of these conceptions overlook the significance of *mutual* vulnerability in engendering trust. Finally, I make a critical turn to Thomas Hobbes, who, despite a number of problematic elements in his political thought, intuitively relies on the notion of mutual vulnerability in his conception of sovereign power and the trust that it confers upon its subjects.

Although a wide variety of factors have been identified that contribute to trustworthy human-robot interaction[5], three of the most commonly

[5] K. Schaefer, J. Chen, J. Szalma, P. Hancock, *A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems*, in "Human Factors: The Journal of the Human Factors and Ergonomics Society", LVIII, 3, 2016, pp. 377-400; P. Hancock, T. Kessler, A. Kaplan, J. Brill, J. Szalma, *Evolving Trust in Robots: Specification Through Sequential and Comparative*

invoked factors center around notions of predictability, behavior, and transparency. For instance, a robot might be considered trustworthy when its actions are predictable, such as by alerting users to impending movements through aural or visual signaling, when it exhibits trust-inducing behaviors, such as by expressing promises to its users[6], or when its decision-making procedures are transparent and do not take place within an impenetrable "black box"[7]. It is worth observing that these notions of trust are not necessarily mutually exclusive, and are often combined to maximize trust in human-robot interaction.

While some research into human-robot interaction has indeed drawn attention to the role of vulnerability in trust[8], it is still a relatively under-researched topic. A useful starting place for understanding the role of vulnerability in trust more generally is the definition given by Denise Rousseau, Sim Sitkin, Ronald Burt, and Colin Camerer, who define trust as "a psychological state comprising the intention to accept vulnerability based upon expectations of the intentions or behavior of another"[9]. Although emphasizing the role of vulnerability in trusting relations, this definition nonetheless offers us little insight into the phenomenon of mutual vulnerability. This is because, according to the authors, only one party must be vulnerable in a trusting relationship; the other party must merely have "intentions or behavior" capable of being expected.

A study on social scientific conceptions of trust carried out by Ann-Marie Nienaber, Marcel Hofeditz, and Philipp Daniel Romeike brings us closer to the relationship between trust and mutual vulnerability. According to the authors, in certain contexts, such as relationships between leaders and followers, trust emerges not when one party accepts a state of vulnerability but when both parties do. This is because trust, according to the authors, depends on "strong emotional based relationships" that require both parties of the relationship to express their vulnerability. To encourage trust, leaders should therefore "avoid showing themselves as distant and inaccessible to their followers. Instead they should demon-

Meta-Analyses, in "Human Factors: The Journal of the Human Factors and Ergonomics Society", 2020, https://doi.org/10.1177/0018720820922080.

[6] L. Cominelli, F. Feri, R. Garofalo, C. Giannetti, M. Meléndez-Jiménez, A. Greco, M. Nardelli, E. Scilingo, O. Kirchkamp, *Promises and Trust in Human-Robot Interaction*, in "Scientific Reports", XI, 1, 2021, https://doi.org/10.1038/s41598-021-88622-9.

[7] W. von Eschenbach, *Transparency and the Black Box Problem: Why We Do Not Trust AI*, in "Philosophy and Technology", 2021, https://doi.org/10.1007/s13347-021-00477-0.

[8] N. Martelaro, V. Nenji, W. Ju, P. Hinds, *Tell Me More: Designing HRI to Encourage More Trust, Disclosure, and Companionship*, in "Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)", 2016.

[9] D. Rousseau, S. Sitkin, R. Burt, C. Camerer, *Not So Different After All: A Cross-Discipline View of Trust*, in "Academy of Management Review", XXIII, 3, 1998, p. 395.

strate their own vulnerability"[10]. Revising the definition given by Rousseau, Sitkin, Burt, and Camerer, we can claim that trust, at least in certain contexts such as relationships between leaders and followers, is a psychological state comprising the intention to accept vulnerability based upon expectations of the vulnerability of another.

Although Nienaber, Hofeditz, and Romeike bring us closer to the role of mutual vulnerability in trust by emphasizing that both parties of the trusting relationship must accept a state of vulnerability, they do not adequately distinguish between notions of individual and mutual vulnerability. That individuals are both vulnerable does not necessarily mean that they are mutually vulnerable. Vulnerable individuals can, for instance, share individual vulnerabilities but to different risks, such as individuals who are corporeally vulnerable but in different risk environments, or even have different individual vulnerabilities entirely, such as individuals from different economic classes. Furthermore, with the distinction between individual and mutual vulnerability in mind, it does not seem intuitive that any expression of vulnerability by a leader will result in greater trust between the leader and her followers. If a leader is vulnerable to a risk that is not shared by her followers (e.g., the risk of her demotion), it would be strange if her expression of it should result in greater trust.

One philosopher who recognized the relationship between mutual vulnerability and trust is Thomas Hobbes. In the *Leviathan*, Hobbes describes what is commonly referred to as a state of nature, a situation in which no governmental power exists to enact and enforce the laws of government. According to Hobbes, individuals in a state of nature are vulnerable to a range of risks and must compete with one another in order to ameliorate them, resulting in a situation "where every man is enemy to every man […] [where] men live without other security, than what their own strength, and their own invention shall furnish them withal"[11]. Although a state of nature is thus marked by the sharing of vulnerabilities such as risks to shelter or nourishment, humans must compete with one another in order to individually ameliorate the risks to which all are vulnerable, ensuring that shared *individual* vulnerabilities (e.g., hunger) never become *mutual* vulnerabilities (e.g., the depletion of shared food reserves).

In order for trust to emerge in a state of nature, individuals must agree to being governed by an all-powerful sovereign. Although Hobbes does not directly refer to notion of mutual vulnerability, its logic is evident in

[10] A. Nienaber, M. Hofeditz, P. Romeike, *Vulnerability and Trust in Leader-Follower Relationships*, in "Personnel Review", XLIV, 4, 2015, p. 577.
[11] T. Hobbes, *Leviathan* (1651), Oxford University Press, Oxford 1998, p. 89.

his understanding of how such a sovereign engenders trust amongst his subjects, such as in the case of a sovereignly enforced covenant:

> If a covenant be made, wherein neither of the parties perform presently, but trust one another; in the condition of mere nature, upon any reasonable suspicion, it is void: but if there be a common power set over them both, with right and force sufficient to compel performance, it is not void.[12]

In other words, the power of the sovereign ensures that failure to uphold a covenant is a risk not only to the promisee, but the promisor as well. Although the vulnerabilities of the promisee and promisor are indeed different, the former being vulnerability with respect to the repercussions of the failed contract and the latter being vulnerability with respect to the expected punishment from the sovereign, they converge on the same risk, namely the failure to uphold the agreed-upon contract.

The trust-building mechanism behind mutual vulnerability, which Hobbes had already intuited – albeit not explicitly formulated – at the time of his writing the *Leviathan*, is relatively straightforward. When two or more individuals are mutually vulnerable, that is, vulnerable to the same risk, they trust that the other(s) will do nothing to put themselves at risk, simply because to put another at risk is to put oneself at risk. Or, in the case of Hobbesian promising, when the promisor puts the promisee at risk by failing to uphold a covenant, she puts herself at risk by exposing herself to the violence of the sovereign's punishment.

Mutual Vulnerability and Autonomy

In addition to promoting trust, mutual vulnerability also has the potential to promote a form of autonomy that I refer to as relational autonomy. In order to develop this concept and its relationship to mutual vulnerability I turn to the thought of Hannah Arendt, who, similar to Hobbes, makes the phenomenon of mutual vulnerability central to her political theory. Unlike Hobbes, however, Arendt locates this phenomenon in a vast array of mundane social interactions, thus bypassing his pessimistic view of human nature and its need for the violent corrective of an all-powerful sovereign.

In *Freedom and Politics*, Arendt criticizes conceptions of freedom that seek to identify it "as one of the inherent attributes of man"[13]. Arendt as-

[12] Ivi, p. 91.
[13] H. Arendt, *Freedom and Politics*, in A. Hunold (a cura di), *Freedom and Serfdom: An Anthology of Western Thought*, Springer, Berlin-Heidelberg 1961, p. 191.

sociates such conceptions of freedom with the "contemplative" tradition of Christianity, particularly as expressed by St. Paul and St. Augustine, who hold the view that "freedom begins when a man withdraws from communal life, from life cheek by jowl with his neighbors – from the sphere, that is, in which the political process is active"[14]. For Arendt, freedom begins in the political sphere, not in the sense that it is through political action that liberties such as a freedom of religion or expression are secured, but in the sense that the political sphere, properly understood, is a domain of interaction.

Arendt's notion of a "space between men" is crucial to her understanding of how human interaction promotes freedom. The significance of this notion for Arendt's conception of freedom is apparent in the final paragraphs of *The Origins of Totalitarianism*, in which she describes totalitarian governments as "destroying all space between men and pressing men against each other"[15], thus depriving them of freedom. Arendt gives a more precise analysis of this "space between men" slightly earlier in *The Origins of Totalitarianism*, where she describes the laws of a constitutional government as the boundaries that protect and support the space necessary for freedom, writing: "To abolish the fences of laws between men – as tyranny does – means to take away man's liberties and destroy freedom as a living political reality; for the space between men as it is hedged in by laws is the living space of freedom"[16].

In *The Human Condition*, Arendt expands on her notion of a "space between men" and introduces the concept of an "in-between" (*Zwischen*) to systematically refer to it[17]. While in *The Origins of Totalitarianism* Arendt focuses on the laws of a constitutional government as constituting the space of this in-between, relatively early in *The Human Condition* Arendt expands the concept to include the shared artifacts of a public world, writing: "To live together in the world means essentially that a world of things is between those who have it common, as a table is located between those who sit around it; the world, like every in-between, relates and separates men at the same time"[18]. Similar to the space constructed by the laws of a constitutional government, Arendt likewise stresses that the space constructed by shared artifacts helps to foster a kind of relational autonomy that simultaneously "relates" and "separates" the humans who share them[19].

[14] Ivi, p. 201.
[15] H. Arendt, *The Origins of Totalitarianism* (1951), Penguin, London 2017, p. 628.
[16] Ivi, p. 611.
[17] H. Arendt, *Vita activa oder vom tätigen Leben* (1960), Piper, München 1994, pp. 52, 173, 192, 199, 237.
[18] H. Arendt, *The Human Condition* (1958), The University of Chicago Press, Chicago 2018, p. 52.
[19] Ivi, p. 53.

Before more closely analyzing how Arendt's concept of an in-between is related to her understanding of freedom, I will turn to a final example: promising. In *The Human Condition*, Arendt most clearly extends the concept to the act of promising when she writes that the "force that keeps [people] together" so as to make them capable of acting in concert and "the power which keeps this public space in existence, is the force of mutual promise or contract"[20]. This passage marks what could be understood as a fundamental shift in Arendt's political thought. While in *The Origins of Totalitarianism* Arendt claims that the laws of a constitutional government are responsible for maintaining the space of human freedom, Arendt in *The Human Condition* claims that promises fulfill this function, which in contrast to laws – in particular those of an all-powerful sovereign – are characterized by temporariness or, as Arendt writes, are temporary "islands of predictability" in which "certain guideposts of reliability are erected"[21].

In order to understand how these different phenomena are examples of what maintains the in-between of human freedom, I will return to the already explicated concept of mutual vulnerability. Although Arendt, like Hobbes, does not explicitly develop any conception of mutual vulnerability, all of her examples of the concept of an in-between can be understood in terms of it, beginning with her understanding of the laws of a constitutional government in *The Origins of Totalitarianism*. To better understand laws in terms of mutual vulnerability, consider those governing torts, that is, any wrongful act that injures or interferes with another's person or property. Interpreted from the perspective of mutual vulnerability, tort laws are guided by the logic of making the perpetrator and victim mutually vulnerable to the risk of the tort being committed. The perpetrator is vulnerable vis-à-vis the punishment she or he must endure should the tort be committed, the victim is vulnerable vis-à-vis the effects of the tort on her or his person or property.

In addition to the laws of a constitutional government, Arendt also identifies shared artifacts as an example of what maintains the in-between that is responsible for human freedom. Although the example that Arendt gives in *The Human Condition* is relatively mundane, namely a "shared table"[22], it is not difficult to furnish other examples of shared artifacts that more clearly display the phenomenon of mutual vulnerability. Public transportation and public goods, such as public infrastructure, help to engender trust in a population by making segments of the population mutually vulnerable to shared risks, namely, the risk of the failure of whatever transportation or infrastructure on which they are dependent.

[20] Ivi, pp. 244-245.
[21] Ivi, p. 244.
[22] Ivi, p. 52.

In countries like the United States or South Africa, the segregation of shared artifacts according to race has likely contributed to the comparatively high degree of social distrust that continues to this day.

A final example that Arendt examines in terms of mutual vulnerability is promising. In *The Human Condition* Arendt describes how the recipient of a promise is not the only party made vulnerable to the possibility of the promise's failure; the giver of the promise is likewise made vulnerable. As Arendt writes, "without being bound to the fulfillment of promises, we would never be able to keep our identities […] which only the light shed over the public realm through the presence of others, who confirm the identity between the one who promises and the one who fulfills," can reveal[23]. While the receiver of the promise might be vulnerable to the promise's failure vis-à-vis the stipulated benefits of the promise, the provider of the promise is vulnerable to the promise's failure vis-à-vis the promise's ability to affirm her or his identity. In other words, for Arendt, failing to uphold our promises would amount to a failure to uphold our identities.

How, though, does mutual vulnerability promote relational autonomy or, as Arendt writes, the phenomenon of simultaneously "relat[ing]" and "separat[ing]" individuals[24]? When individuals are mutually vulnerable to shared risk and consequently trust that those who are likewise vulnerable will do nothing to put themselves at risk, they grant each other greater freedom to act in ways that might be unpredictable or unexpected, knowing that ultimately they will avoid the shared risk in spite of any momentary periods of unpredictability or unexpectedness. In this way, a form of spontaneity is encouraged through mutual vulnerability that lies at the core of Arendt's conception of freedom, which, as Maurizio Passerin d'Entrèves describes, is neither "simply the ability to choose among a set of possible alternatives" nor "the faculty of *liberum arbitrium* […] given to us by God" but rather "the capacity to begin, to start something new, to do the unexpected, which all human beings are endowed by virtue of being born"[25].

To better understand the relationship between mutual vulnerability and the spontaneity of relational autonomy, I will return to the example of being mutually vulnerable to a joint work project. When two colleagues are vulnerable to the risk of their project's failure, they trust that neither will do anything to cause the project's failure, because to do so would not only put the other at risk but oneself at risk. Consequently,

[23] Ivi, p. 237.
[24] Ivi, p. 53.
[25] P. D'Entrèves, *The Political Philosophy of Hannah Arendt*, Routledge, London 1994, p. 66.

when either colleague violates any subsidiary norms guiding the completion of the project (e.g., failing to adhere to the deadline of a draft or missing a day of work), they are confident that these momentary lapses do not indicate that the colleague will tolerate a global failure of the project itself. It is this trust that consequently encourages the tolerance of a certain degree of unpredictability or unexpectedness within the context of the working relationship and thus the experience of a greater degree of individual autonomy for both colleagues.

Finally, it is worth briefly discussing the fundamental difference between the conceptions of mutual vulnerability that motivate the political theories of Hobbes and Arendt. While it is by now evident that both rely on some notion of mutual vulnerability to engender interpersonal trust and, for Arendt, relational autonomy, their understandings of the phenomenon are vastly different. While Arendt sees the concept as characterizing a vast array of social phenomena, from the sharing of artifacts to the making of promises, Hobbes locates the concept in the singular social phenomenon of contracts violently enforced by an all-powerful sovereign. Arendt would have thus likely viewed a state of nature devoid of any form of mutual vulnerability and composed, as Christine DiStefano writes, "of a body politic of social orphans who have socially acculturated themselves" as an absurd figment of a Hobbesian imagination[26].

To summarize, Arendt's concept of an "in-between" or "space between men" is best understood as a space of mutual vulnerability. By establishing mutual vulnerability to a shared risk, not only does greater interpersonal trust emerge but a form of relational autonomy as well, in which the trusting parties grant each other the possibility to act in ways that are spontaneous, that is, unpredictable and unexpected.

Mutually vulnerable human-robot interaction: challenges and possibilities

A significant degree of research has investigated the potential effectiveness of designing affective and, by extension vulnerable, robots. Rosanne M. Siino, Justin Chung, and Pamela J. Hinds have shown that robots that use affective language to disclose information are generally liked better by their users[27]. Although this study does not focus on the relationship

[26] C. Di Stefano, *Configurations of Masculinity: A Feminist Perspective on Modern Political Theory*, Cornell University Press, Cornell 1991, p. 92.
[27] R. Siino, J. Chung, P. Hinds, *Colleague vs. Tool: Effects of Disclosure in Human-Robot Collaboration*, proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication, 2008.

between affective expression and trust specifically, a more recent study by Nikolas Martelaro, Victoria C. Nneji, Wendy Ju and Pamela Hinds has shown that expressions of vulnerability by the robot generally have a positive impact on human trust[28]. Furthermore, recent research by Sara Strohkorb Sebo, Margaret Traeger, Malte Jung, and Brian Scassellati has shown that expressions of vulnerability by a robot in team settings help to increase trust amongst human team-members, in what has been called a "ripple effect" of empathy extending from human-robot interaction to human-human interaction[29].

Although research shows that expressions of vulnerability likely do improve human trust in robots and that expressions of mutual vulnerability would therefore also likely improve trust, the mere expression of vulnerability is not sufficient for the emergence of relational autonomy. This is due to the fact that crucial to the phenomenon of relational autonomy is the experience of being given greater freedom to act in ways that are unexpected or unpredictable. For this to occur, the robot would have to be able to first effectively perceive the vulnerability of the human and, having determined that the human is indeed mutually vulnerable, then attune the degree of freedom it grants the human based on its perception. As previously mentioned, such "attunements" might take the form of relaxing restrictions on subsidiary deadlines of a joint project, knowing that adherence to the final deadline presents a risk to which the human is vulnerable and thus a risk that the human will still avoid, despite potentially forgoing subsidiary deadlines.

The difficulty of engineering robots capable of both perceiving mutual vulnerability and attuning its behavior in response to this perception is perhaps already clear. Because mutual vulnerability is dependent on sharing risk as opposed to vulnerability, in order for the robot to attune the freedom that it grants to a human it must have a high degree of context-dependent insight into the human's life. For instance, in order for relational autonomy to emerge out of mutual vulnerability in the context of a joint human-robot work project, the robot would need to perceive the human's vulnerability to the risk of the project's failure. This vulnerability is dependent upon a vast array of potentially fluctuating factors. A worker who is already successful, for example, might be less vulnerable to the risk of a project's failure than a less successful worker. In a similar

[28] N. Martelaro, V. Nenji, W. Ju, P. Hinds, *Tell Me More: Designing HRI to Encourage More Trust, Disclosure, and Companionship*, proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2016, p. 184.
[29] S. Sebo, M. Traeger, M. Jung, B. Scassellati, *The Ripple Effects of Vulnerability: The Effects of a Robot's Vulnerable Behavior on Trust in Human-Robot Teams*, proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, 2018.

vein, a worker whose values conflict with those of the project would likewise be considered less vulnerable to the risk of the project's failure than a worker whose values do in fact align.

The problem of perceiving vulnerability becomes particularly evident in the mutual vulnerability that characterizes relationships of companionship. To better understand how mutual vulnerability manifests itself in companionship, it is helpful to turn to the conception of self-developed by psychologists associated with the Stone Center at Wellesley College, such as Judith Jordan[30]. In *Feminist Morality*, Virginia Held summarizes the findings of Jordan, writing that she conceptualizes the self not as being rigidly individualistic but as

> having both a need for recognition and a need to understand the other […] [in which] both give and take in a way that not only contributes to the satisfaction of their needs as individuals but that affirms the 'larger relational unit' they compose. Maintaining this larger relational unit then becomes a goal.[31]

In contrast to the mutual vulnerability that characterizes relationships such as contract agreements, which renders the related individuals mutually vulnerable vis-à-vis a good extrinsic to the relationship (e.g., the timely performance of the contract), the mutual vulnerability that characterizes relationships composed of a "larger relational unit" is intrinsic to the relationship itself. That is, the related individuals become mutually vulnerable to the risk of the dissolution of the relationship.

Held cites the "mother-child relation" as a paradigm of the kind of relationship that derives its value intrinsically[32]. Contrasting it with relations that derive their value egoistically, such as that of a contract agreement, Held writes that the "emotional satisfaction of a person engaged in mothering arises from the well-being and happiness of another human being and from the health of the relation between the two persons"[33]. Because both mother and child are vulnerable to the risk of the dissolution of the relationship, mother-child relations are consequently characterized by a high degree of "permanence"[34], as Held writes, and frequently persist in spite of fundamental changes that occur at the individual level. This permanence in turn demonstrates the relational autonomy that characterizes most healthy parent-child relations, that is, the autonomy

[30] J. Jordan, *The Meaning of Mutuality*, in "Work in Progress", XXIII, 1986, pp. 1-11.
[31] V. Held, *Feminist Morality*, The University of Chicago Press, Chicago 1993, p. 60.
[32] Ivi, p. 204.
[33] Ivi, p. 205.
[34] Ivi, p. 206.

that emerges when individuals trust that their relationship will continue in spite of any individual changes they might personally undergo.

Although Held analyzes the intrinsic value of relationships with respect to the paradigm of the mother-child relationship, a variety of relationships can be said to derive their value intrinsically. Friendships, partnerships, and various forms of companionship more generally are often characterized by this experience of mutual vulnerability. Long-term friends, for instance, will likely be confident that both are equally committed to the friendship and that both would be equally disappointed should the relationship dissolve. Momentarily lapses in observing the norms that govern friendships, such as the forgetting of a date, will consequently likely be more liberally handled, implying that long-term friendships lend a degree of autonomy that short-term acquaintances, for example, do not. In short, the more committed a relationship is, the more it has the potential to confer greater relational autonomy to its individual parties.

If the phenomenon of relational autonomy indeed extends to companion relationships, then it would be particularly difficult to design relationally autonomous human-robot companion relationships. This is because the indications that signal another's vulnerability to the health of the relationship itself, and not to any extrinsic benefit that the relationship might bring, are even more nuanced and variegated than those that signal another's vulnerability to, for instance, the success of a joint work project. It often takes considerable psychological insight to be able to determine when an individual values a relationship for its intrinsic as opposed to extrinsic benefits. Furthermore, in order to arrive at the point of being capable of perceiving a human's mutual vulnerability to the health of their relationship, the robot would first have to be capable of fostering a relationship capable of being intrinsically valued.

To conclude, the difficulties of replicating the form of relational autonomy that emerges from the mutual vulnerability of human relations in human-robot interaction are myriad. Instead of attempting to design a form of mutual vulnerability into human-robot interaction sufficient to allow for the emergence of relational autonomy, a more efficient approach to ensuring that users of companion robots continue to experience benefits of relational autonomy is to ensure that they continue to have meaningful human relations, in spite of whatever social responsibilities their companion robot might gain. A simple solution would be to ensure that individuals assigned companion robots are also assigned a social worker to assist in the robot's utilization, which would consequently allow for the possibility of relational autonomy emerging between social worker and robot user.

Furthermore, when a social worker is assigned to assist a user in their utilization of their robot, the possibility arises that mutual vulnerability

manifests itself not only with respect to the health of their relationship, but with respect to the success of their joint project, namely the successful utilization of the companion robot. When both social worker and robot user are invested in the success of the robot and avoidant of its failure, the recognition of this mutual vulnerability would lead the social worker to grant the robot user greater liberty in creatively utilizing their robot. Inversely, the robot user would likewise grant greater liberty to the social worker by, for example, entertaining suggestions for utilizing the robot in ways that might at first seem unintuitive.

6. Concluding Remarks

This paper has used the phenomenon of mutual vulnerability, namely the situation of two or more individuals being vulnerable to the same risk, as a point of departure for analyzing human-robot interaction. After having defined mutual vulnerability, I argued that the phenomenon is responsible not only for promoting interpersonal trust, but also for promoting a form of relational autonomy. Mutual vulnerability promotes trust due to the fact that when two or more individuals are vulnerable to the same risk, they trust that no one will purposefully put the other at risk. Furthermore, mutual vulnerability promotes autonomy due to the fact that when two or more individuals perceive that they are vulnerable to the same risk, they grant each other greater freedom to act in ways that may be unpredictable or unexpected, confident that no one, in spite of their spontaneous behavior, will purposefully put the others at risk.

Although it would indeed be feasible to program robots that are capable of expressing mutual vulnerability and, as recent research suggests, would even be likely that this would promote trust in human-robot interaction, I have nonetheless claimed that the mere expression of mutual vulnerability is not sufficient for the promotion of relational autonomy. In order for relational autonomy to emerge, not only must both individuals be mutually vulnerable, they also must be capable of perceiving the vulnerability of the other and adjusting the freedom they grant each other in accordance with their perception. Given the complexity of this task, I have concluded that it is presently unrealistic to attempt to design robots capable of promoting relational autonomy. Instead, I have proposed that the state-sanctioned implementation of social robots such as companion robots should include the employment of social workers, through whom the emergence of mutual vulnerability and relational autonomy can be made possible.

Finally, although mutual vulnerability is likely significant to our experiences of both trust and autonomy in human interaction, this is not to say

that neither trust nor autonomy can be achieved in our relations with robots whatsoever. Some individuals have in fact reported a greater sense of autonomy when interacting with robots, knowing that they will not be negatively judged[35]. Others have pointed out that in some care settings receivers of care should *not* be granted greater autonomy, due to the nature of their needs and limitations[36]. Furthermore, as already mentioned, various degrees of trust are already being engineered into human-robot interaction. Whether a lack of mutual vulnerability and the related phenomenon of relational autonomy necessarily means a reduction in interaction quality can ultimately only be determined on a case-by-case basis, in which the unique and individual needs of humans are taken into account.

Bibliography

Arendt H., *The Origins of Totalitarianism* (1951), Penguin, London 2017.
Id., *The Human Condition* (1958), The University of Chicago Press, Chicago 2018.
Id., *Vita activa oder vom tätigen Leben* (1960), Piper, München 1994.
Id., *Freedom and Politics*, in A. Hunold (a cura di), *Freedom and Serfdom: An Anthology of Western Thought*, Springer, Berlin-Heidelberg 1961.
Coeckelbergh M., *Care Robots and the Future of ICT-Mediated Elderly Care: A Response to Doom Scenarios*, in "AI and Society", a. XXXI, n. 4, 2016.
Id., *How to Describe and Evaluate "Deception" Phenomena: Recasting the Metaphysics, Ethics, and Politics of ICTs in Terms of Magic and Performance and Taking a Relational and Narrative Turn*, in "Ethics and Information Technology", a. XX, n. 1, 2018.
Cominelli L., Feri, F., Garofalo, R., Giannetti, C., Meléndez-Jiménez, M., Greco, A., Nardelli, M., Scilingo, E., Kirchkamp, O., *Promises and Trust in Human-Robot Interaction*, in "Scientific Reports", a. XI, n. 1, 2021, https://doi.org/10.1038/ s41598-021-88622-9.
d'Entrèves P., *The Political Philosophy of Hannah Arendt*, Routledge, London 1994.
Di Stefano C., *Configurations of Masculinity: A Feminist Perspective on Modern Political Theory,* Cornell University Press, Cornell 1991.
Engelhart K., *What Robots Can–and Can't–Do for the Old and Lonely*, in "The New Yorker", May 24, 2021, https://www.newyorker.com/magazine/2021/05/31/what- robots-can-and-cant-do-for-the-old-and-lonely.
Hancock P., Kessler T., Kaplan A., Brill J., Szalma J., *Evolving Trust in Robots: Specification Through Sequential and Comparative Meta-Analyses*, in "Human

[35] S. Turkle, *Alone Together: Why We Expect More From Technology and Less from Each Other*, Basic Books, New York 2011, p. 10.
[36] M. Coeckelbergh, *Care Robots and the Future of ICT-Mediated Elderly Care: A Response to Doom Scenarios*, in "AI and Society", XXXI, 4, 2016, p. 460.

Factors: The Journal of the Human Factors and Ergonomics Society", 2020, https:// doi.org/10.1177/0018720820922080.
Held V., *Feminist Morality*, The University of Chicago Press, Chicago 1993.
Hobbes T., *Leviathan* (1651), Oxford University Press, Oxford 1998.
Jordan J., *The Meaning of Mutuality*, in "Work in Progress", a. XXIII, 1986.
Mackenzie C., Stoljar N. (a cura di), *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self*, Oxford University Press, Oxford 2000.
Martelaro N., Nenji V., Ju W., Hinds P., *Tell Me More: Designing HRI to Encourage More Trust, Disclosure, and Companionship*, in "Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)", 2016.
Nienaber A., Hofeditz M., Romeike P., *Vulnerability and Trust in Leader-Follower Relationships*, in "Personnel Review", a. XLIV, n. 4, 2015.
Rousseau D., Sitkin S., Burt R., Camerer C., *Not So Different After All: A Cross-Discipline View of Trust*, in "Academy of Management Review", a. XXIII, n. 3, 1998.
Schaefer K., Chen J., Szalma J., Hancock P., *A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems*, in "Human Factors: The Journal of the Human Factors and Ergonomics Society", a. LVIII, n. 3, 2016.
Sebo S., Traeger M., Jung M., Scassellati B., *The Ripple Effects of Vulnerability: The Effects of a Robot's Vulnerable Behavior on Trust in Human-Robot Teams*, in "Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction", 2018.
Sharkey A., Sharkey N., *We Need to Talk About Deception in Social Robotics!*, in "Ethics and Information Technology", 2020, https://doi.org/10.1007/s10676-020-09573-9.
Siino R., Chung J., Hinds P., *Colleague vs. Tool: Effects of Disclosure in Human-Robot Collaboration*, in "Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication", 2008.
Sparrow R., Sparrow L., *In the Hands of Machines? The Future of Aged Care*, in "Minds and Machines", a. XVI, n. 2, 2006.
Turkle S., *Alone Together: Why We Expect More From Technology and Less From Each Other,* Basic Books, New York 2011.
von Eschenbach W., *Transparency and the Black Box Problem: Why We Do Not Trust AI*, in "Philosophy and Technology", 2021, https://doi.org/10.1007/s13347-021-00477-0.