HUMAN-MACHINE COMMUNICATION

Human-Machine Communication
Volume 9, 2024
https://doi.org/10.30658/hmc.9.5

Human-Machine Relations and Relational Autonomy

Hendrik Kempt[1], Camilla Francesca Colombo[1], and Saskia Kathi Nagel[1]

1 RWTH Aachen University, Applied Ethics Group, 52062 Aachen, Germany

Abstract

The ever-growing use of interactive technologies and digitization processes in decision-making contexts raises key challenges to the conceptualization of autonomy. Starting from the shortcomings of individualistic models, we argue that relational accounts of autonomy provide a more adequate characterization of human beings as socially embedded and integrated. By introducing a decision-theoretic framework, and further theoretical elaboration, we suggest a substantial account of relational autonomy incorporating machines and technologies within the significant relations entertained by agents. Besides optimizing decision-making performances, human-machine interactions could genuinely support and enhance individual autonomy, by fostering human connections, expanding the individual's social context with additional perspectives, and prompting self-reflection. We conclude that, with crucial caveats, while we should not automate our relations, relational autonomy can benefit from automation and digitization elsewhere.

Keywords: relational autonomy, autonomy support, human-machine interaction, decision theory

Funding Acknowledgment: This research was partly funded by the project VEREINT (grant number 16SV9108), and the project DecEnt (grant number 01GP2210A) funded by the German Federal Ministry for Research and Education (BMBF).

CONTACT Hendrik Kempt • hendrik.kempt@humtec.rwth-aachen.de • RWTH Aachen University • Applied Ethics Group • Theaterplatz 14 • 52062 Aachen • Germany

ISSN 2638-602X (print)/ISSN 2638-6038 (online)
www.hmcjournal.com

Introduction

Most humans deem autonomy to be a highly valuable good, and a precondition for a free and democratic society with emancipated people. Ensuring that members of a society can make well-informed decisions based on authentic preferences and desires is a prerequisite for free discourse and collective decision-making based on democratic procedural means. With the ever growing mechanization and digitization of our decision-making contexts, people's ability to assert their autonomy has become both expanded and strained at the same time: on the one side, we never had more information and thus opportunity to form authentic preferences and make decisions based on those preferences; on the other side, the leading role of technology in people's lives has made them vulnerable to undue manipulation in those very decisions by sludges, dark patterns, fake news, and more.

Relational autonomy accepts the inevitability of external influences not only upon our decision-making but indeed upon our lived experience as humans (cf. Lyreskog et al., 2020), and aims to conceptualize conditions and constraints under which humans can be considered autonomous within their social connections and integration. This framework allows for the analysis and explanation of different forms of socially conditioned autonomy. It does not center the isolated individual but persons as socially embedded beings, and it offers suggestions on how to improve autonomy by recognizing, supporting, and improving their social relationships.

However, at this stage, the role of technology for and in relational autonomy has been reflected upon only to an insufficient degree. The incorporation of interactive and personalized technologies in our everyday lives poses a particularly crucial challenge that any account of autonomy, including a relational one, ought to answer. This article aims to first provide a conceptual answer to the question whether interactive and autonomous technology can become embedded in our relational contexts in a way that it is genuinely contributing to our autonomy, and second to evaluate whether such a possibility would be desirable.

We claim it is possible to consider autonomous machines as potentially part of the network of social relations that ensures our autonomy, even when they are at odds with our human relations. While we caution to equate these contributions to those of human relationships (as machines lack fundamental capabilities akin to humans), they can indeed improve our decisions in many different ways. Especially when our social network of relations is not enabling our full autonomy but is oppressive or dysfunctional, a machine that is geared toward providing a reflective moment for decision-makers can substantially improve our autonomous decision-making. How this is possible within the frame of relational autonomy, however, is dependent on how one reconstructs relational autonomy, human-machine collaboration, and what counts as decisional autonomy.

The structure to make our argument is thus the following. To build a common ground about the concepts in question, we elaborate on the key notions. We first characterize the relational autonomy account and describe how it amounts to a distinct way of understanding the decision-making complexities of embedded social beings. Second, we turn toward identifying the human-machine interactions our argument will target. As there is a large variety of interactions discussed in philosophical discourse, with varying degrees of interactive complexity and influence on decision-making, it is necessary to elaborate on the specific interactions we concentrate on. Third, in turning toward decision theory, we lay the groundwork for assessing the consequences of increased influence of autonomous machines in our decisional processes. Fourth, we consider the impact human-machine relations can have on relational autonomy and traditional models of decision theory. In this, we argue for the thesis that we should not automate relational autonomy; however, relational autonomy can benefit from automation and digitization elsewhere.

Autonomy as a Relational Concept

The idea that autonomy ought to be primarily understood throughout the relationships an agent entertains in her social context is the key element of influential models in ethics and political theory, largely due to Mackenzie and Stoljar's seminal work (2000). Although this idea can be found in other and earlier communitarian philosophies, such as the Ubuntu philosophy (Menkiti, 1984; Samkange, 1980), and has been successfully incorporated in feminist accounts (such as Meyers, 1989), Mackenzie and Stoljar were pivotal in reconceptualizing a structured account of relational autonomy within the given framework of (Western) philosophy. By now, this account has become a crucial feature of political theory and feminist philosophy, which acknowledges, centers, and grounds subjects in their relations to others.

Previous accounts of autonomy have traditionally analyzed this concept mainly referring to the capacity of individuals to make their own decisions, often with a certain claim against the (undue) influence of others (Frankfurt, 1971). As such, autonomy is conceptualized as a capacity expressed by agents in their own self-determination. This default position, with the emergence of relational accounts, has been labeled as (hyper)individualistic (cf. Christman, 2004; Nagel, 2015). This individualistic interpretation does not acknowledge social relations as a condition for autonomy, but rather as a consequence of it: within this view, we can establish meaningful and stable relationships with others because we are autonomous beings, with fully formed preferences and capacities.

Contrary to this approach, relational autonomy focuses on the supporting and enabling role social relations can play in individuals' formation of a self-concept, their desires, and, crucially, in their decision-making processes (Niker et al., 2021). As socially embedded beings, we are formed and informed by our social relations, we rely on them to navigate the world, and make decisions that transform our own lives. By extending these assumptions to other agents, relational autonomy demands the recognition and respect of others in their specific social contexts, as some human choices cannot be truly understood from a fully isolated and individualized perspective. In this sense, however, relational autonomy puts the established notion of the authenticity of preferences and desires into question, leading to critical implications.

The focus on authenticity not as a hierarchical first-order/second-order understanding of preferences (cf. Dworkin, 1988; Frankfurt, 1971), but as an expression of learned identities (even in potentially harmful or oppressive contexts) offers an alternative solution to the traditional issue of the tension between first-order and second-order desires. While Frankfurt considers first-order desires as misguided or inauthentic, if they are not endorsed by second-order reflective (and, thus, more sophisticated) procedures, authors like Friedman (1986) and more recently Mackenzie (2019) leave room for a rival interpretation: first-order desires are actually more authentic to the preferences of a person, and second-order reflective desires might be the ones considered inauthentic when conflicting with first-order preferences.

Similarly, Diana Meyers's account of autonomy competency (1989) provides some theoretical space for a relational idea of competency: without considering the *social scaffolding* that is necessary to develop the competency to critically reflect on the social norms one has internalized, this person cannot be considered (fully) autonomous. Thus, in order to *enhance* someone's autonomy, not only the cognitive, but also the emotional and social development of a person have to be accounted for.

Such accounts, however, usually rely upon a strong substantive idea of autonomy (i.e., they purport that the preferences we form ought not to be content-neutral; although Meyers, 1989 remains neutral on that point). If, for example, an agent has internalized some harmful stereotype and formed their preferences on this basis (these are often referred to as *adaptive preferences*), she cannot, again, be considered (fully) autonomous. Within this interpretation, oppressed people in oppressive systems, while making *their own decisions*, cannot be considered fully autonomous, as they have adapted their preferences to fit the autonomy-inhibiting context of their decisions—even when their second-order desires are settled and their cognitive abilities established. The danger of reducing autonomy to the cognitive capacity to find the right means to realize one's primary and secondary desires (without reflecting on the social conditions in which they form) lies in missing out on some fundamental improvements of autonomy at scale: If a society creates cognitively autonomous, but emotionally and socially subjugated and oppressed people, every individual might appear autonomous, without nonetheless resulting in a society of autonomous agents.

All these issues translate to a different assessment of practices and to demands for autonomy *support,* stressing the role of relations (Mackenzie & Stoljar, 2000; Nagel, 2015; Nagel & Reiner, 2013): In viewing humans, their preferences, agency, and decision-making processes as intimately intertwined with the world that shaped them, we can gain a clearer view on how each person can be supported to make more authentic decisions.

Relational autonomy has been gaining more and more influence in bioethics and medical ethics, since the decision-making capacities and autonomy of agents in highly vulnerable states such as sickness is often resting on trusted relationships with others (see analyses in Baumann, 2008; Mackenzie & Stoljar, 2000; Nagel, 2015; Oshana, 2016; Sherwin, 1998). This account has even become a standard approach to questions of autonomy in end-of-life-, palliative care-, and critical care-decisions (Gómez-Vírseda et al., 2020; Grignoli et al., 2018; Wilson et al., 2014), where the "the vulnerabilities and insecurities of patients require a perspective that goes beyond questions about the right to choose without coercion or deception" (Nagel, 2015, p. 49).

All these features and desiderata make relational autonomy a fruitful account for analyzing and assessing how and to which degree a genuinely autonomous agent's decision-making capacities can be expanded and supported. This is particularly the case for all contexts in which individuals can be construed as socially embedded (e.g., for digitized complex decision-making processes).

Human-Machine Collaborations

That human-machine interactions grow, both in quantity and in quality, is self-evident at the present stage. The mechanization and digitization of communication, public discourse, economic and social activities, many work tasks, and even elements of our intimate lives have created a near-permanent exposure to, and demand for, interaction with smart and autonomous machines of different kinds, for different purposes, and with different intensity and attention requirements. Before we can discuss the impact of our interactions with these machines within the relational autonomy framework, we should however distinguish among the possible interactive instances we face in using autonomous machines. Specifically, while some of these interactions are individual, others take place in contexts where technology is ubiquitous or even *environmental* (see, e.g., Aydin et al., 2019; Valera, 2019).

We interact with machines on many different levels and with a variety of purposes. Some machines we use as a medium to communicate with other human beings, like smartphones and the social media apps therein (in forms of computer-*mediated* communication). Some others we use as tools for purposes outside these machines, like mobility, home-making, information retrieval, shopping, and so on. And some machines we use to achieve purposes that are confined within these machines, such as entertainment, distraction, some quasi- or para-social interactions (e.g., chatting with a chatbot), and so forth. All these interactions occur on an individual level with machines that agents perceive as tools over which they exert some control.

On a larger scale, we interact with complex systems that may comprise a variety of machines and humans that rely on machines in their decision-making processes. These interactions often are necessary, as we cannot *not* interact with these quasi-environmental systems. In these instances, the perception of interactivity is less one of a discursive purpose-realization but rather a near-constant re-alignment with the requirements of the technology in a larger purpose-driven action-context (take, e.g., the interactions of a worker with a highly complex industrial machine).

Notably, philosophical discourse on interactions between humans and machines deals largely with the first dimension: human-machine collaborations (Nyholm, 2018; Simmler & Frischknecht, 2021), human-machine partnerships, robot coworkers, and other characterizations of human-machine interactions all presuppose an individuated machine rather than a larger system with whom humans interact and to which they react.

On the other hand, if the analysis focuses on technical systems, such as AI at large (rather than individuated chatbots or other AI-powered devices), it is less about *interaction* in the sense explained above but rather about the interaction between human societies and their norms on the one side, and technology on the other.

These different forms of interactions—ranging from individual micro-interactions with a machine to achieve specific goals to a vague, ubiquitous environment—are influencing our abilities to make decisions, and thus our autonomy, in a fundamental way. As we are investigating whether machines could be considered part of our network of relations constituting and ensuring our autonomy, we will focus on the individuated machines and our interactions with them. This includes the potential of using interactive technology both as a medium to connect with others and the consequences of such potential for relational autonomy, as well as interacting with autonomous technologies itself in a human-agent interactive instance (HAI) and others (Bradshaw et al., 2011). There are other relevant accounts for HMC scholars that conceptualize the role of technology for our autonomy (e.g., Agent-Network Theory; Latour, 2005) or Self-Determination Theory (Deci & Ryan, 1987), but their contribution cannot be discussed in sufficient detail in this paper.

In the following section, we will move from the human-machine collaboration debate to introduce a conceptual framework which will help highlight how both human relationship and interactive technologies or digitization processes can shape the decision-making context.

Decision Theory and Autonomy

The concept of autonomy has been variously construed in moral philosophy, political philosophy, and bioethics as the capacity of an individual to *act* and make one's *decisions* in accordance with specific criteria and desiderata. These criteria can be broadly carved up as *internal* and *external* conditions for autonomy. While the former refer to an agent capacity "to be one's own person, to be directed by considerations, desires, conditions, and characteristics that are (. . .) part of what can somehow be considered one's authentic self" (Christman & Anderson, 2005, p. 23), the latter require that an agent's decisions "are not the product of manipulative or distorting external forces" (Christman, 2004, p. 154), thus stressing freedom from undue influence. As discussed in Section 2, this characterization of individual autonomy suffers from some shortcomings the relational framework aims to address. On the one hand, internal conditions of autonomy require from the agent sophisticated reasoning skills and self-reflection abilities including second-order coherence and consistency such that most, if not all, individuals would fall short in meeting these standards. This limit of individualistic accounts is even more prominent in complex choice scenarios calling for quick responses or existential decisions, like in medical settings (Nagel, 2015). On the other hand, external conditions presuppose an individualistic dimension of choice, which has been criticized as unrealistic and untenable within a more comprehensive characterization of individuals as intrinsically socially embedded (see Baumann, 2008; Niker et al., 2021; Oshana, 2016), who can experience "cognitive transformative experiences (. . .) which are also shared, interpersonal, and deeply relational processes" (Lyreskog et al., 2020, p. 85). Nonetheless, for proponents of relational autonomy accounts, one critical issue remains the adequate identification of which forms and kinds of external outputs are legitimate, or even welcomed, without infringing upon one's autonomy.

In the present section, we examine a possible account of relational autonomy which includes relations with machines using the tools of Decision Theory. This theoretical framework offers some useful insights to appreciate how an individual's deliberative skills, critical reflection and judgment capacities, and decision-making processes can be enhanced and promoted by the support of trusted and expert human and nonhuman relations. At the same time, this reconstruction starts sketching adequate boundaries and limitations for a legitimate relation with machines and digitization processes in general which is autonomy preserving.

Decision Theory is one of the most influential frameworks employed in the social sciences to model choice behavior. The key elements of a standard decision-theoretic model are the following (Bradley, 2017):

- An **agent**, who selects among a set of available **actions** a_1, a_2, . . . a_n;
- The **states** s_1, s_2, . . ., s_n, the descriptions of events which may impact, or be relevant to, the output of the decision;
- The **outcomes** $o_{1,1}$, $o_{1,2}$, . . ., o_{nn}, all the possible combinations of actions and states.

These elements, defining a **decision problem**, are usually represented by a decision matrix:

FIGURE 1 Example of Decision Matrix

	s_1	s_2	s_3	. . .
a_1	$o_{1,1}$	$o_{1,2}$	$o_{1,3}$	. . .
a_2	$o_{2,1}$	$o_{2,2}$	$o_{2,3}$	. . .
. . .	. . .	. . .	. . .	. . .

These further elements are then introduced:

- P = a **probability** function defined over the set of states, capturing the agent's expectations and beliefs (a distribution of probabilities over S).
- U = a **utility** function, representing the agent's preferences over the outcomes (an ordering function over O). Utilities can be thus thought of as the agent's values informing that decision.
- A **decision criterion**, which is usually the maximization of expected utility, but could also be maximin, maximax, and so forth, depending on the agent's risk attitudes.

A decision is then defined as the selection of a specific action from the set by the agent, performed in accordance with a given decision criterion.

Let us now examine more closely what these different elements modeling choice behavior aim to capture by the means of a toy decision problem. Suppose you are offered to subscribe to a health insurance plan for the next 10 years, either premium or basic, or to continue without health insurance at all. Of course, this decision will depend on whether you will have health issues in the following years. All the combinations of your available actions (premium insurance plan, basic health insurance plan, no insurance plan), and of the states of nature (health problems, no health problems), can be represented in the following decision matrix:

FIGURE 2 Decision Matrix for the Choice Among Different Health Insurance Plans

	Health Problems (20%)	No Health Problems (80%)
Premium health insurance	6	3
Basic health insurance	4	5
No health insurance	0	10

The numbers in each cell of the matrix are traditionally characterized as a representation of the agent's desires, goals, and personal tastes: in this sense, and with all the obvious limitations, utilities map the agent's own preferences and feelings toward the possible consequences of the decision problem she is facing. In this example, the agent is scared of being sick without a health insurance, she is quite bugged by the idea of paying for the premium package if she will be healthy, but, having pursued the basic package, she is then almost indifferent between a situation where she uses it because she is sick (4) and a situation where her money goes to waste but she is healthy (5). The probability function over the state space, according to Jeffrey's (1965) most influential interpretation, expresses subjective probability judgments: in short, 20% and 80% stand for the agent's own beliefs and expectations over her future health conditions. Eventually, the decision criterion employed by the agents expresses her attitude toward risk-taking: maximin (e.g., choosing in a way which minimizes the worst potential outcome) amounts to an exceptionally pessimistic and risk-averse approach to decision-making, and maximax—maximizing the best potential outcome—lays on the other extreme of the risk-taking spectrum. We can think of our toy agent as a fairly moderate and reasonable individual, using the standard decision criterion endorsed by cost-benefit analysis, the maximization of expected utility, where every outcome is weighed for its likelihood, and then the expected utility of all the actions is computed:

$$U_{(\text{premiuminsurance})} = (6 \times 0.2) + (3 \times 0.8) = 3.6$$
$$U_{(\text{basicinsurance})} = (4 \times 0.2) + (5 \times 0.8) = 4.8$$
$$U_{(\text{noinsurance})} = (0 \times 0.2) + (10 \times 0.8) = 8$$

The decision of getting no insurance plan (the action with the highest expected utility), in this scenario, is the one which should express the nuanced and complex feelings, attitudes, beliefs, and values of the agent, with respect to all these different dimensions of choice. Arguably, when an agent builds and fills in the decision matrix herself, using her own personal utilities, subjective probability judgments, and risk-attitudes, she is autonomous in taking that decision, respecting both the internal and external conditions traditionally required by individual autonomy (see Colombo & Nagel, 2023).

This reconstruction, however, is far too simplistic. First, choice problems are not faced by the agent in a void. Even when the decision-maker does not ask for advice or support directly, her utilities and subjective probability judgments are influenced, to say the least, by the context she lives in, her personal experiences, and significant relations. From a descriptive point of view, arguing that a decision is autonomous only insofar as the decision matrix is solely build by the agent, and filled with values and beliefs which are uniquely and distinctively her own, would mean to overlook the whole preferences and beliefs formation processes, which are crucial to self-reflection and personal identity. A relational account of autonomy, in this sense, amounts to a more naturalistic representation of the actual preferences and beliefs formation processes (see e.g., Christman, 2004). As Nagel argues, an adequate and full understanding of individual autonomy in choice contexts "should include the examination of effects of relational ties on decision making to gain insights on how different influences can strengthen, limit, or be ambiguous with respect to people's decision-making desires and competences" (2015, p. 51).

From a normative point of view, the individualistic understanding of decision-making is not even the most efficient, reasonable, nor adequate approach to most choice problems. In our health plan case, the agent could benefit from expert opinion on insurance practices and medical advice, which could provide relevant information she is not aware of (as in the case of a genetic consultant) or shaping the probability functions in a way that more closely maps the current empirical data. At the same time, trusted and long-lasting family or friend opinions could lead to a more comprehensive appreciation of one's attitude toward risk-taking. But this is not even the most crucial aspect in which decision-making processes could be improved by a relational approach.

Starting from the 1970s, behavioral economists, cognitive psychologists, and neuroscientists have started investigating choice behavior using the tools and techniques proper of each specific discipline. What these different lines of research have in common is their challenge to the standard models of human behavior grounded in concepts of instrumental or economic rationality. Far from being idealized rational agents, meeting strict coherence and consistency requirements, human beings are prone to all sorts of biases, reasoning flaws, and mistakes (Gigerenzer, 2015; Kahneman et al., 2021). An extensive analysis of biases goes of course beyond the scope of this paper, but our health insurance example can provide a handy exemplification. One common bias our agent can incur when assigning utilities is the so-called temporal discount of future utilities. In behavioral economics, cognitive psychology, and neuroeconomics this phenomenon is vastly documented and robust, and it is defined as the tendency to give greater value to rewards as they move away from their temporal horizons and toward the "now" (Frederick et al., 2002, p. 353). In this case, people will tend to assign a higher utility to the money they have to spend now to buy insurance, and to assign a lower utility to the better health conditions and health care they might expect in the future. Another very pervasive bias agents could experience when facing this kind of decision problems is the general difficulty in making sense of probabilities and risks, especially when very low percentages are involved. Extensive empirical studies (such as Levy & Baron, 2005) have shown that people's understanding and actual grasp of the relative figures of small risks is often limited or blatantly defected. This pervasive difficulty is particularly pronounced in the medical field (Mourali & Yang, 2023) where multiple risks add up and the chances of most diseases occurring are usually small, but can differ dramatically in their proportions.

The decision matrix the agent builds, and the choice she eventually makes, within an individualistic autonomy account, is in this sense very likely to be prone to these reasoning flaws and biases. She will probably underestimate her future health problems if she is in good shape now; she will neglect future health care expenditures, while overvaluing the money health care packages will cost her in the present. Eventually, she might incur in miscalculations of expected utilities, which are also very well-documented in decision-makers (List & Haigh, 2005). The final outcome of this choice, in short, would not be the most *reasonable*, and possibly not even the closest to the individual's *real* preferences and beliefs, as the decision-making process is not efficiently and authentically representing these characteristics in the decision matrix.

A relational account of autonomy, where other agents and machines collaborate with the individual in modeling and solving the choice problem, could thus be beneficial to the decision-maker, addressing these cognitive biases and enhancing reasoning capacities. In our toy case, medical personnel could help in focusing on some salient scenarios, and so could do personal testimonies from friends and family members sharing trusted and solid relations, overcoming the discounting of future utility bias. Technological devices could also be part of this broader net of meaningful and trusted interactions, when adequately shaped: graphic depictions of magnitudes of risks, visual aids, and interactive interfaces have been proven effective in improving humans' understanding and grasp of percentages and probabilities (Fraenkel et al., 2018). Boosting techniques based on machine learning algorithms could be as well employed in various settings to enhance decision performances and task-solution skills and have been tested in choice contexts as fruitful AI-generated decision aids (Becker et al., 2022). To be clear, research on algorithm driven decision-making also advises caution as various forms of so-called algorithmic biases can be conveyed by the use of technological decisional aids. As Kordzadeh & Ghasemaghaei (2022) argue, however, extensive and conclusive empirical studies on this potential danger are still lacking, and there is no reason to think that algorithmic biases could not be adequately addressed and balanced by human design and oversight. Therefore, we take in this paper this optimistic stance, and conclude that all these external interventions and contributions to the decision-making process can be construed as autonomy-supporting rather than autonomy-infringing, insofar as the resulting decision matrix, and the final outcome of the decision, amounts to a more authentic and well-reasoned representation of her underlying values and beliefs.

In all these examples, the significant human-machine interactions can be modeled as helping the agent in building and filling in the decision matrix, rather than altering it by taking out options which were previously available in the actions set, or by changing the utility payoffs directly using incentive/disincentive schemes. These crucial differences with cases of explicit nudging or manipulation provide significant indications as to how to conceptualize an adequate relational autonomy account of human-machine collaboration.

Toward a Substantial Account of Relational Autonomy in Human-Machine Relations

In the following, we will discuss how the concept of relational autonomy fits with interactive technology, especially when considering decision-making contexts. To this end, we first look at the benefits and at the role interactive technologies can play in contributing to our autonomy. This groundwork is especially compelling given our analysis in Section 3: while technology is creating ever more complex and embedded instances of interaction and has led to some practical accounts of how to incorporate autonomous machines in the autonomy of human agents (cf. De Visser et al., 2018), we observe in the literature a comparative lack of theoretical work on this matter (a recent exception here is Mhlambi & Tiribelli, 2023).

To start, interactive technology can improve our decision-making skills by cognitively enhancing our understanding of the different available options, their potential consequences, or even of the world at large. More specifically, as we describe in Section 3, autonomous machines can help overcome widespread biases, and thus optimize decision-making performances. While we appreciate these benefits, which are also the kind of cognitive support the proponents of individualist accounts of autonomy focus on, we also contend that they exhaust the complex role interactive technology plays in shaping the concept of autonomy. In this sense, we argue that interacting with machines does not only foster our cognitive abilities, but can also strengthen our relational autonomy by enriching our emotional and social capacities, as well as our relations (even, or maybe especially, when our social relations are insufficiently doing so).

Take, for example, a young pregnant woman living in a highly conservative social context. Suppose she is inclined to get an abortion, as she is neither in the emotional nor financial state to take care of the baby she would have. She cannot talk to her family members and even to most of her friends—many of them devoted mothers—about this, fearing rejection and disdain even at the thought of terminating the pregnancy. In consulting an interactive chatbot that allows her to reflect on her preferences, make financial plans, offers information about abortion clinics, safety measures, and other information she could experience a richer supporting environment for making the right decision for her.

Such a *reflective moment*, enabled through interactive technology, can equip us with adequate tools for appreciating the differences between our deeply held preferences and some of the (dubious) moral judgments or expectations of many around us. In this sense, then, some human-machine interactions can enhance the relational autonomy of their users.

From this perspective, interactive technology can be fruitfully incorporated even in substantial accounts of relational autonomy. Not only can we reduce biases in our decision performances, especially when those biases permeate our social relations and amount to learned secondary preferences (i.e., our assessments and perceptions of risk); but it can also uncover the deeply oppressive or harmful context in which we grow up and provide a *rational* and less biased influence in our decision-making processes.

Further, the near-permanent availability of technologies might result in forming bonds which were previously unavailable, and thus expand our relations, and also in improving our decision-making abilities by connecting with people close to us (some evidence for this effect can be found in, e.g., Quan-Haase et al., 2019). Being able to maintain and deepen connections independent of our physical presence has been and will further be a fundamental shift in how we conduct relationships and has become a new standard of mutual expectations in our communicative effort and availability. Especially during the COVID-19 pandemic, interactive technological means, often with autonomous elements, were key to maintaining many social relationships, and this has crucially challenged the philosophical assessment of technology (Gómez-Virseda & Usanos, 2021).

However, these undeniable benefits provided by interactive and autonomous technology ought to be more critically analyzed from a philosophical-conceptual perspective. A substantial account of relational autonomy should address the conceptual role played by machines and technologies, and not be simply reduced to a list of potential advantages and risks of human-machine interactions. In this sense, it seems indisputable that machines are principally limited in establishing those relationships that are contributing to relational autonomy. The richness and depth of human-human relationships cannot be emulated by current (and foreseeable) technological developments (for a more optimistic view on this possibility, see Coeckelbergh, 2012 or Gunkel, 2023). If a machine is reacting to their user as an individual, it is not doing so in appreciating that user's personality *individually*, but rather by mapping its training data onto the user's input and matching with other similar responses. Even highly interactive, autonomous technologies do not possess a coherent concept of the human interactant as a person, but rather correlate the person's specific preferences with a general picture of standardized and categorized (i.e., profiled) preferences. The personalizing features—name, special interests, and the like—are usually provided by the user and thus not genuine judgments about the user's character and estimations about what they—as relating individuals—might need. Any advice or suggestion from autonomous interactive technologies—from medical preferences to musical playlists—is an extrapolation based on one's own and *other* users' preferences, and not a judgment on someone's character or personality.

These limits are also evident in the normative consequences and implications of autonomous technologies' relations to human users. The inability to blame them for mistakes, their lack of genuine agency, own interests, and thus trustworthiness clearly point to the fact that their contribution to an agent's relational autonomy is severely limited philosophically. Machines, in this sense, might not be as *relatable* as they would be required to, in a substantial relational autonomy sense. This lack of relatability is also critical for the conceptualization of our preference-formation process, which—within a relational account—is informed and dependent on inputs and mutual responses with our social ties and context. If we consider human-machine relations as adding substantially to our relational autonomy, we have to explain how adding autonomous technology to this social context is a genuinely positive element in the procedural conditions of individuals forming beliefs and making decisions.

The benefits of human-machine interaction for relational autonomy, together with the philosophical-conceptual reservations we should nonetheless keep, seem to stand in conflict with each other. However, we hold these claims to be perfectly compatible once the scope of the influence of interactive technology is clarified. While the replacement of sources of autonomy cannot and should not be the goal, we should not merely understand interactive technologies as contributing to our cognitive abilities: autonomous and interactive machines can improve and expand our social relationships, support those we already have, and enhance our autonomy by doing so.

Thus, we should not automate our relations when seeking autonomy; however, relational autonomy can benefit from automation and digitization elsewhere.

Conclusion

In this paper, we argue that autonomous machines and interactive technologies can be incorporated among the significant relations entertained by an autonomous agent—within a relational autonomy framework. Our proposal not only captures the pervasive trend of human-machine collaborations, but also aims to provide a substantial account of how non-human interactions can fit in and might even enhance individuals' relational autonomy. We argue that this framework could offer a novel and fruitful theoretical tool for examining the scope of specific instances of human-machine interactions and their impact on individual autonomy, building on the *reflective moment* insight of what makes digitized choice settings meaningful and autonomy supporting. On the one hand, our decision-making performances could benefit from the use of interactive technologies, which can help overcome biases and support the building and shaping of the choice problem itself, eventually achieving *better* decisions. However, while human-machine interactions can be beneficial to decision-making in a relational account, this does not imply that machines can or should be conceptualized as being able to form meaningful relationships with humans.

Moreover, the use of technologies and digitization processes could exacerbate the critical problem of distinguishing between legitimate autonomy support and undue influence: automating decision-making processes, while making them more efficient, could also conceal and neglect individuals' own values, personal attitudes, and approaches to choice. Within a decision theoretic perspective, this would mean substituting the agent in building and solving the decision matrix, by profiling and other automatic task-solution techniques. More significantly, and substantially, machines could be construed as merely reacting to human's inputs and generalizing rather than addressing the agent individually. In this sense, they would not reach the level of richness and personalization accomplished by human relations. This worry is well-grounded, and we certainly do not argue here that human-machine collaboration can adequately replace human interactions in our decision-making practices. Nonetheless, technologies can help foster and maintain human connections, provide further perspectives on a choice situation, thus expanding the individual's social context, and offer key inputs for self-reflection and critical evaluation of one's own values, beliefs, and decisional processes. For all these reasons, and with some caveats, human-machine interactions can be conceptualized in a relational account as genuinely autonomy preserving and potentially even enhancing. While in the present paper we only take the first steps in this effort, the features of substantively autonomy-supporting interactions highlighted here can provide the groundwork for a more systematic and comprehensive analysis of practical decision settings and technologies, leading to the development of a full-fledged model of technologically-mediated reflective moments in decision-making.

Author Biographies

Hendrik Kempt (PhD, RWTH Aachen University) is a Postdoctoral Researcher at the Applied Ethics Group at RWTH Aachen University. His research is mainly on ethical implications of emerging human-machine interactions, foundational models, and digitized societies. He has published several books on these topics, including on the possibility of human-machine friendship (2022) and (un-)explainable technologies (2024).

https://orcid.org/0000-0002-5886-2987

Camilla F. Colombo (PhD, London School of Economics) is a Postdoctoral Researcher at the Applied Ethics Group at RWTH Aaachen University. She is a moral philosopher and decision theorist; her research focuses on the cognitive and theoretical foundations of decision theory and on decision-making in ethically sensitive choice scenarios. Her PhD thesis was on the causal and moral relevance of the doing/allowing distinction.

https://orcid.org/0000-0002-6190-8931

Saskia K. Nagel (PhD, University of Osnabrück) is full professor and Chair of the Applied Ethics Group at the RWTH Aachen University (Germany). Her research ranges from ethics of technology, ethics of AI and human-technology relations, to bioethics and the ethics of neuroenhancement, as well as responsibility and trust in increasingly digitized societies.

https://orcid.org/0000-0001-9657-5121

Acknowledgments

We thank Niel Conradie, Frieder Bögner, Giacomo Figa-Talamanca, Niklas Michels, Michelle Görlitz, and Henry Seidel as well as two anonymous reviewers for their helpful input and assistance on earlier drafts of this paper.

References

Aydin, C., Woge, M. G., & Verbeek, P.-P. (2019). Technological environmentality: Conceptualizing technology as a mediating milieu. *Philosophy and Technology*, *32*(2), 321–338. https://doi.org/10.1007/s13347-018-0309-3

Baumann, H. (2008). Reconsidering relational autonomy. Personal autonomy for socially embedded and temporally extended selves. *Analyse & Kritik*, *30*(2), 445–468. https://doi.org/10.1515/auk-2008-0206

Becker, F., Skirzyński, J., van Opheusden, B., & Lieder, F. (2022). Boosting human decision-making with AI-generated decision aids. *Computational Brain & Behavior*, *5*(4), 467–490. https://doi.org/10.1007/s42113-022-00149-y

Boy, G. A. (2017). *The handbook of human-machine interaction: A human-centered design approach*. CRC Press.

Bradley, R. (2017). *Decision theory with a human face*. Cambridge University Press.

Bradshaw, J., Feltovich, P. J., & Johnson, M. (2011). Human-agent interaction. *Handbook of Human-Machine Interaction*.

Christman, J. (2004). Relational autonomy, liberal individualism, and the social constitution of selves. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *117*(1/2), 143–164.

Christman, J., & Anderson, J. (2005). *Autonomy and the challenges to liberalism: New essays*. Cambridge University Press.

Coeckelbergh, M. (2012). *Growing moral relations. Critique of moral status ascription*. Springer.

Colombo, C. F., & Nagel, S. K. (2023). A decision-theoretic approach to assisted medical decision-making. *AJOB Neuroscience*, *14*(3), 241–243. https://doi.org/10.1080/21507740.2023.2243866

Deci, E. L., & Ryan, R. M. (1987). The support of autonomy and the control of behavior. *Journal of Personality and Social Psychology, 53*(6), 1024–1037.

Denier, Y., & Gastmans, C. (2022). Relational autonomy, vulnerability and embodied dignity as normative foundations of dignified dementia care. *Journal of Medical Ethics*, *48*(12), 968–969. https://doi.org/10.1136/jme-2022-108722

De Visser, E. J., Pak, R., & Shaw, T. H. (2018). From 'automation' to 'autonomy': The importance of trust repair in human–machine interaction. *Ergonomics, 61*(10), 1409–1427.

Dworkin, G. (1988). *The theory and practice of autonomy (Cambridge studies in philosophy).* Cambridge University Press. https://doi.org/10.1017/CBO9780511625206

Fraenkel, L., Reyna, V., Cozmuta, R., Cornell, D., Nolte, J., & Wilhelms, E. (2018). Do visual aids influence patients' risk perceptions for rare and very rare risks? *Patient Education and Counseling*, *101*(11), 1900–1905. https://doi.org/10.1016/j.pec.2018.06.007

Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy*, *68*(1), 5–20. https://doi.org/10.2307/2024717

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, *40*(2), 351–401. https://doi.org/10.1257/002205102320161311

Friedman, M. A. (1986). Autonomy and the split-level self. *Southern Journal of Philosophy*, *24*(1), 19–35. https://doi.org/10.1111/j.2041-6962.1986.tb00434.x

Gigerenzer, G. (2015). *Simply rational: Decision making in the real world*. OUP.

Gómez-Vírseda, C., de Maeseneer, Y., Gastmans, C. (2020). Relational autonomy in end-of-life care ethics: A contextualized approach to real-life complexities. *BMC Journal of Medical Ethics* (1), 50. https://doi.org/10.1186/s12910-020-00495-1

Gómez-Vírseda, C., & Usanos, R. A. (2021). Relational autonomy: Lessons from COVID-19 and twentieth-century philosophy. *Medicine, Health Care, and Philosophy*, *24*(4), 493–505. https://doi.org/10.1007/s11019-021-10035-2

Grignoli, N., Di Bernardo, V., & Malacrida, R. (2018). New perspectives on substituted relational autonomy for shared decision-making in critical care. *Critical Care*, *22*(1), 260. https://doi.org/10.1186/s13054-018-2187-6

Gunkel, D. (2023). *Person, thing, robot*. MIT Press.

Jeffrey, R. C. (1965). *The logic of decision*. University of Chicago Press.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Little, Brown Spark.

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems, 31*(3), 388–409.

Latour, B. (2005). *Reassembling the social: An introduction to actor-network theory*. Oxford University Press.

Levy, A. G., & Baron, J. (2005). How bad is a 10% chance of losing a toe? Judgments of probabilistic conditions by doctors and laypeople. *Memory & Cognition*, *33*(8), 1399–1406. https://doi.org/10.3758/BF03193372

List, J., & Haigh, M. (2005). A simple test of expected utility theory using professional traders. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 945–948. https://doi.org/10.1073/pnas.0408022101

Lyreskog, D. M., Karlawish, J., & Nagel, S. K. (2020). Where do you end, and I begin? How relationships confound advance directives in the care of persons living with dementia. *The American Journal of Bioethics*, *20*(8), 83–85. https://doi.org/10.1080/15265161.2020.1781967

Mackenzie, C. (2019). Relational autonomy: State of the art debate. In A. Armstrong, K. Green, & A. Sangiacomo (Eds.), *Spinoza and relational autonomy: Being with others* (pp. 10–32). EUP.

Mackenzie, C., & Stoljar, N. (2000). *Relational autonomy: Feminist perspectives on autonomy, agency, and the social self*. Oxford University Press

Menkiti, I. A. (1984). Person and community in African traditional thought. In R. Wright (Ed.), *African philosophy, an introduction*. University Press of America.

Meyers, D. T. (1989). *Self, society, and personal choice*. Columbia.

Mhlambi, S., & Tiribelli, S. (2023). Decolonizing AI ethics: Relational autonomy as a means to counter AI harms. *Topoi*, *42*(3), 867–880. https://doi.org/10.1007/s11245-022-09874-2

Mourali, M., & Yang, Z. (2023). Misperception of multiple risks in medical decision-making. *Journal of Consumer Research*, *50*(1), 25–47. https://doi.org/10.1093/jcr/ucac040

Nagel, S. K. (2015). When aid is a good thing: Trusting relationships as autonomy support in health care settings. *The American Journal of Bioethics*, *15*(10), 49–51. https://doi.org/10.1080/15265161.2015.1074316

Nagel, S. K., & Reiner, P. B. (2013). Autonomy support to foster individuals' flourishing. *American Journal of Bioethics*, *13*(6), 36–37. https://doi.org/10.1080/15265161.2013.781708

Niker, F., Felsen, G., Nagel, S. K., & Reiner, P. B. (2021). Autonomy, evidence-responsiveness, and the ethics of influence. In M. J. Blitz & J. C. Bublitz (Eds.), *The law and ethics of freedom of thought, Volume 1: Neuroscience, autonomy, and individual rights* (pp. 183–212). Springer International Publishing. https://doi.org/10.1007/978-3-030-84494-3_6

Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human-robot collaborations and responsibility-loci. *Science and Engineering Ethics*, *24*(4), 1201–1219. https://doi.org/10.1007/s11948-017-9943-x

Oie, S. (2023). From relational freedom to autonomy: An expansion of Verbeek's postphenomenology. *Human Studies*, *46*, 1–20. https://doi.org/10.1007/s10746-023-09681-7

Oshana, M. (2016). *Personal autonomy in society*. Routledge.

Quan-Haase, A., Zhang, R., Wellman, B., & Wang, H. (2019). Older adults on digital media in a networked society: Enhancing and updating social connections. In M. Graham & W. H. Dutton (Eds.), *Society and the internet: How networks of information and communication are changing our lives*. Oxford. https://doi.org/10.1093/oso/9780198843498.003.0006

Samkange, S. J. T. (1980). *Hunhuism or ubuntuism: A Zimbabwe indigenous political philosophy*. Graham.

Sherwin, S. (1998). *The politics of women's health: Exploring agency and autonomy*. Temple University Press.

Simmler, M., & Frischknecht, R. (2021). A taxonomy of human–machine collaboration: Capturing automation and technical autonomy. *AI & SOCIETY*, *36*(1), 239–250. https://doi.org/10.1007/s00146-020-01004-z

Valera, L. (2019). New technologies. Rethinking ethics and the environment. In L. Valera & J. C. Castilla (Eds.), *Global changes: Ethics, politics and environment in the contemporary technological world* (pp. 29–43). Springer International Publishing. https://doi.org/10.1007/978-3-030-29443-4_4

Wilson, F., Ingleton, C., Gott, M., & Gardiner, C. (2014). Autonomy and choice in palliative care: Time for a new model? *Journal of Advanced Nursing*, *70*(5), 1020–1029. https://doi.org/10.1111/jan.12267

HUMAN-MACHINE

COMMUNICATION