HUMAN-MACHINE COMMUNICATION

Human-Machine Communication
Volume 9, 2024
https://doi.org/10.30658/hmc.9.3

Four Notions of Autonomy. Pitfalls of Conceptual Pluralism in Contemporary Debates

Roman Wagner[1] and **Bert Heinrichs**[2,3]

1 University of Bonn, German Reference Centre for Ethics in the Life Sciences (DRZE), Bonner Talweg 57, 53113 Bonn, Germany
2 University of Bonn, Institute of Science and Ethics (IWE), Bonner Talweg 57, 53113 Bonn, Germany
3 Forschungszentrum Jülich, Institute of Neurosciences and Medicine: Brain and Behaviour (INM-7), Wilhelm-Johnen-Straße, 52428 Jülich, Germany

Abstract

The concept of autonomy is indispensable in the history of Western thought. At least that's how it seems to us nowadays. However, the notion has not always had the outstanding significance that we ascribe to it today and its exact meaning has also changed considerably over time. In this paper, we want to shed light on different understandings of autonomy and clearly distinguish them from each other. Our main aim is to contribute to conceptual clarity in (interdisciplinary) discourses and to point out possible pitfalls of conceptual pluralism.

Keywords: autonomy, second-order desires, self-legislation, artificial systems, conceptual pluralism

Introduction

The concept of autonomy is indispensable in the history of Western thought. At least that's how it seems to us nowadays. However, the notion has not always had the outstanding significance that we ascribe to it today. And its exact meaning has also changed considerably over time. In ancient Greece, autonomy was primarily a political category (cf. Pohlmann,

CONTACT Roman Wagner • wagner@drze.de • University of Bonn • German Reference Centre for Ethics in the Life Sciences (DRZE) • Bonner Talweg 57 • 53113 Bonn • Germany

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

1971). Autonomy referred to the independence of a city-state from another power and the ability to govern its own political affairs. It was only occasionally used in ethical contexts. A prominent example of this latter use is when Sophocles describes Antigone's attitude as autonomous (Sophocles, 1891, *Antigone*, line 821). With the Romans and in the Latin Middle Ages, the concept of autonomy played only a marginal role. It was not until modern times (i.e., from the 17th century onward) that the concept regained importance, initially primarily in the law (cf. Schneewind, 1997). Eventually, Immanuel Kant turned the concept of autonomy into a key philosophical concept, and the meaning he gave to it continues to be highly relevant.

While the notion of autonomy predates Kant's philosophy, the concept of self-determination, understood as self-*legislation*, achieved in his writings a systematic significance and sophistication unmatched before, serving as a constant point of reference for all subsequent debates on the subject. That, of course, does not mean that there were no important precursors to Kant's specific understanding of autonomy—most notably Jean-Jacques Rousseau. Still, it is no exaggeration to say that in Kant's philosophy, the idea of autonomy took on a form that in many ways shaped the course of philosophical thought, both in agreement as well as in disagreement. The Kantian idea of a kind of freedom which only rational and self-conscious beings can manifest was accepted in form, if not in all details, by the subsequent thinkers of German Idealism, most notably by Johann Gottlieb Fichte (cf. Allison, 2013). To this tradition, Kant's understanding of autonomy presented a counterpoint to the rationalistic determinism found, for example, in Spinoza and promised a way to understand ourselves as morally responsible beings that are, in at least some ways, fundamentally different from anything else to be found in nature.

For many, at least the general outline of Kant's understanding of autonomy remains authoritative. Others, however, have proposed alternative readings, some of which deviate considerably from Kant. This variety of accounts of autonomy sometimes lead to serious misunderstandings, especially in interdisciplinary debates. The situation has become more acute since robots and computer systems have gained ever greater degrees of independence from human control and are therefore often referred to as *autonomous systems*. This use of the term is hardly compatible with the Kantian understanding of autonomy, at least given the current state of technological development. In fact, it is unclear whether robots and computers will ever be autonomous in the Kantian sense. In this paper, we will not speculate about the prospects of whether artificial systems will ever be autonomous. Rather, we want to shed light on different understandings of autonomy and clearly distinguish them from each other. Our main aim is to contribute to conceptual clarity in (interdisciplinary) discourses and to point out possible pitfalls of conceptual pluralism.

There are a great number of philosophical accounts of autonomy, and we cannot consider them all here. We have limited ourselves to four of these approaches that we believe are particularly important for the current debate. In addition to Kant's, these approaches are the ones developed by Harry Frankfurt, Tom Beauchamp and James Childress, and Lucian Floridi. While in Kant's philosophy, autonomy is tied to morality, Harry Frankfurt's hierarchical or reflexive approach is one of the most influential nonmoral—though not completely nonnormative—philosophical conceptions of autonomy. For the discussion of applied ethics, the nonhierarchical, nonreflexive, and nonmoral understanding of autonomy by Tom Beauchamp and James Childress is of great importance. We will therefore discuss it in some detail. Last, in Lucian Floridi's notion of autonomy, we will consider a recent variant that is particularly relevant to discussions of artificial intelligence (AI).

While the four approaches to autonomy mentioned are by no means the only ones to be found in philosophical discourse (cf. the diverse contributions in Christman, 2014 and for a social-relational account cf. Oshana, 2006), they are arguably the most influential and widespread ones. To understand at least some of the basic ideas that are tied up in the concept of autonomy, it will be helpful to compare and evaluate them more closely. In order to do that, we will not develop them historically, but systematically, starting with the conceptually sophisticated approaches of Frankfurt and Kant, followed by the conceptions of autonomy that are most influential in applied ethics, and finally discussing Floridi's approach.

It will become clear that the Kantian, as well as the Frankfurtian, concept of autonomy are intimately tied to an understanding of what kinds of beings we humans are. This understanding, in turn, is the basis for the moral assessment of autonomy. In contrast, a morally neutral account of autonomy allows to include complex systems acting independently of human control. However, only if we determine what autonomy means can we assess whether it is valuable and worthy of protection.

Finally, we will discuss whether it might be better to abandon the notion of autonomy in order to avoid misunderstandings. We will argue that conceptual pluralism can be useful but has its limits when it leads to serious misunderstandings and misconceptions. We advocate a deliberate use of the term autonomy that always discloses which understanding of autonomy is being adopted. In this way, we take account of the fact that philosophy does not on its own decide on the correct or incorrect use of terms in nonphilosophical debates. However, philosophy can point to different traditions and ways of using terms, work out implications, and thus make a contribution to the interdisciplinary debate on autonomy.

Four Notions of Autonomy

Harry Frankfurt: The Hierarchical Approach

One of the most influential and widely received approaches to autonomy in 20th-century philosophy is that of Harry Frankfurt. To Frankfurt, what is indicative of autonomy is at the same time what is exclusively indicative of the kind of being that we humans are: persons. Persons, according to Frankfurt, are beings whose will takes on a specific form. They differ from other animals insofar as they can form not only desires to act in a certain way but also desires of a higher order (i.e., second-order desires or volitions), which have first-order desires as their objects.

A person who is driven by a compulsive desire to act in a certain way, without reflectively endorsing this compulsion, like an unwilling addict, cannot, according to Frankfurt, be considered as acting autonomously. Such a person acts freely since her actions can be explained by reference to her own first-order desires—the addict wants to take the drug—but she does not act with free will. Only in cases where the person acts according to those first-order desires she reflectively endorses, can her action be understood as the result of free will, hence autonomous. A person who is driven to act against his own reflective evaluation, is, in Frankfurt's words, "estranged from himself" and "a helpless or passive bystander to the forces that move him" (Frankfurt, 1971, p. 17).

Autonomy, in Frankfurt's approach, therefore takes on a reflective form. A person cannot count as being autonomous unless she is able to distance herself from her first-order desires and evaluate them from a higher perspective. In his early writings, Frankfurt explains that "the capacity for reflective self-evaluation (. . .) is manifested in the formation of second-order desires" (1971, p. 7). Note that this understanding of autonomy goes beyond the idea that free action is the expression of a mere choice based on a desire. For a being to count as acting autonomously, it has to be conscious of its own first-order desires and it needs to take an evaluative stance toward these desires. In Frankfurt's understanding, persons are, therefore, necessarily self-conscious and able to reflect on what they *should do*. In this weak sense, Frankfurt's conception of autonomy is normative. However, its standard remains purely subjective. Whatever the higher-order desires of a person are, they serve as normative criteria of practical reflection upon which autonomously chosen actions rest.

There is an objection that was raised early on against Frankfurt's theory of freedom and autonomy. This criticism points to a dilemma with which Frankfurt's approach seems to be confronted. Either the will is constituted by an infinite chain of higher-order desires, or this chain is cut off arbitrarily. Both horns of the dilemma are untenable. However, Frankfurt anticipated this objection and reacted to it with the idea of a decisive commitment:

> When a person identifies himself *decisively* with one of his first-order desires, this commitment 'resounds' throughout the potentially endless array of higher orders. (. . .) The fact that his second-order volition to be moved by this desire is a decisive one means that there is no room for questions concerning the pertinence of desires or volitions of higher orders. (Frankfurt, 1971, p. 16)

Although he repeatedly refined and clarified his initial approach in later writing (some of which are collected in Frankfurt, 1995), the basic idea remained the same throughout later publications. In the series of desires, there are some basic desires which are expressive of who a person *really* is. These desires are taken to be subjectively necessary, which is to say that the person cannot but identify with them. Whenever a person acts in opposition to these desires, she is only a passive and nonautonomous bystander of her own actions. These desires express a person's true or authentic self. More than that, a person constitutes herself by wholeheartedly identifying with certain desires:

> When the decision is made without reservation, the commitment it entails is decisive. (. . .) The decision determines what the person really wants by making the desire on which he decides fully his own. To this extent the person, in making a decision by which he identifies with a desire, constitutes himself. The pertinent desire is no longer in any way external to him. (Frankfurt, 1995, p. 161)

The standard of an autonomous choice is, therefore, the decisive identification with a certain subset of desires. Although this conception is normative (if only in a weak sense), it is decidedly not moral, for morality typically claims some form of objective validity. The normativity Frankfurt refers to is exclusively based on our subjective interests and inclinations:

> What we care about has to do with our particular interests and inclinations. If what we *should* care about depends upon what we do care about, any answer to the normative question must be derived from considerations that are manifestly subjective. (. . .) Answers to the normative question are certainly up to us in the sense that they depend upon what we care about. However, what we care about is not always up to us. Our will is not invariably subject to our will. We cannot have, simply for the asking, whatever will we want. There are some things that we cannot help caring about. (Frankfurt, 2006, pp. 24–25)

Frankfurt's idea of autonomy comes with high demands. Only self-conscious beings, for whom some things are particularly important, can be autonomous. As far as we know, only humans fulfill these conditions. Animals, on the other hand, lack self-consciousness, as do artificial systems. Moreover, the latter have no things to care about in an appropriate way.

One major problem with this theory of autonomy is that it is not clear what a decisive commitment exactly is. No matter how committed a person is at any point in time, it seems always possible that she might have a change of heart. Wholehearted commitment can therefore never guarantee that a person takes something she believed to be a constitutive part of herself to be something she now understands to be external to her. Frankfurt's attempt to solve this problem by employing the concept of love and volitional necessities (cf. Frankfurt, 2006, p. 32ff) (i.e., the idea of things we cannot but care about given our nature) points in the direction of an Aristotelian form of normativity grounded in a normative understanding of our life form but has never been systematically developed by Frankfurt himself (cf. for such an approach Foot, 2003). As it stands, it is not entirely clear how this approach can be squared with Frankfurt's commitment to volitional subjectivism and how he can avoid the charge of deriving normative statements from statements of psychological fact.

Another problem concerns the question of the value of autonomy. Although there is something to the idea that acting autonomously in the way Frankfurt envisions has moral value, it seems also clear that morally horrendous actions can be autonomous in Frankfurt's sense. If a person's higher-order volitions or volitional necessities commit this person to morally reprehensible acts, their autonomy is obviously at odds with morality. While Frankfurt recognizes and addresses this problem in his later work (cf. Frankfurt, 2006, p. 46f), it is hard to see how, given the conceptual tools Frankfurt has developed and employed in his understanding of a free will, he can make sense of the moral dimension of autonomy. Even if we accept Frankfurt's hierarchical theory of autonomy as conceptually compelling, it is far from clear that autonomy understood this way has any moral value (cf. O'Neill, 2003a, pp. 5–6).

Despite these criticisms, Frankfurt's concept of autonomy is without doubt one of the most influential in contemporary philosophy. When people talk about autonomy in philosophical contexts today, it is not unlikely that they are referring to consistent second-order volitions in Frankfurt's sense.

Immanuel Kant: Autonomy as Self-Legislation

Kant's conception of autonomy exhibits many parallels to Frankfurt's approach, at least superficially. Like Frankfurt, Kant takes autonomy to be indicative of what persons are. Like Frankfurt, Kant believes that autonomy requires self-consciousness. Like Frankfurt, Kant thinks that autonomy expresses the freedom of the will. Like Frankfurt, Kant does not believe that the power of choice explains autonomy. A closer look, however, reveals that these similarities are, in fact, points of opposition. After all, Kant has a special view of what self-consciousness amounts to as well as of how self-consciousness is intertwined with practical reason and the will.

To Kant, the will, self-consciousness, autonomy, and practical reason are not just connected—they are different facets of one and the same complex phenomenon. Self-consciousness, according to Kant, is *rational self-consciousness*. Hence, self-consciousness includes theoretical and practical reason. Beings like us or persons are able to determine their own actions (i.e., to act autonomously), precisely because of their ability to self-consciously determine their own actions according to the representation of a law, as Kant notes, when he writes:

> Everything in nature acts according to laws. Only a reasonable being has the power to act according to the representation of the laws, that is to say according to principles; in other words, he has a will. Will is nothing but practical reason. (Kant, 1785, AA IV.412)

In acting according to the representation of rational laws, persons exemplify a unique kind of freedom. This freedom is autonomy and in acting autonomously, persons have a spontaneous, non-receptive knowledge of these principles. Unlike Frankfurt, Kant does not think that the act of wholehearted identification with any desire or volition expresses the standpoint of an acting person. Rather, the person determines her standpoint through practical reason in a way that is objectively valid. In other words, persons do not discover their standpoint by reflecting upon those desires they wholeheartedly identify with. Instead, they determine their standpoint by applying universally valid principles of practical reason to their individual situation and asking, "How should I act?" They answer this question by forming a subjective maxim, which they judge to be either (formally) valid or not.

Only if a subjective maxim has a specific form, namely the form of a categorical imperative, can it be considered as internal to the will. This means that no hypothetical imperative can be the fundamental principle of practical reason. Where hypothetical imperatives require externally given ends (i.e., contingent desires), a categorical imperative is a valid principle of action, no matter what a person contingently wants. As Stephen Engstrom explains: "[Kant claims] that practical reason's most basic imperatives, those of morality, are categorical rather than hypothetical in form. Human reason must accordingly be conceived as autonomous, as the sole source of its principles of action" (Engstrom, 2009, p. ix).

Both Frankfurt and Kant agree that at least a certain form of desire is the ground for autonomous action. Both also agree that these grounds are normative. However, there are two important aspects of Kant's thought that differ fundamentally from Frankfurt's account:

First, no psychological state can ever determine autonomous actions. Second, the kind of causality that explains human action requires rational self-consciousness.

According to Kant's theory, a person should act on a maxim that has the right form, a form expressed by the Categorical Imperative. Because the Categorical Imperative, as the internal principle of reason, determines her action, the action can count as an expression of autonomy. Nothing external to practical thought determines what the person does. In this way, she is free of the external influence of inclination (i.e., merely arbitrary psychological determinations of the will). Because of this, Kant's conception neither suffers from a looming infinite regress nor dogmatism. There is no regress because the Categorical Imperative is the supreme principle of action, not just one desire in a series of desires. There is no dogmatism since the Categorical Imperative is not just any principle, but rather the constitutive principle of practical reason and therefore the constitutive principle of the will. The principle of practical reason, in Kant's approach, is the standard not only of autonomy but of morality (i.e., the supreme normative principle of *how we should act* as rational beings). There is no difference between the will, practical reason, and morality. Because of this close connection, the moral value of Kantian autonomy is not in question. Autonomous action simply *is* morally justified action and therefore autonomous action has moral value.

Although it was developed more than 200 years ago, Kant's understanding of autonomy is also very present in contemporary philosophy and at the same time the subject of ongoing interpretative controversies, which we have of course not been able to trace here. However, it should have become clear that Kant significantly raises the requirements for the attribution of autonomy in comparison to Frankfurt: Only persons as rational beings can act autonomously. As natural beings, we humans by no means always act autonomously, but as noumenal beings, we are at least capable of doing so in principle.

Tom Beauchamp and James Childress: Intentionality, Understanding, Non-Control

In sharp contrast to these very sophisticated approaches to autonomy are those that have been developed in the field of applied ethics in recent decades. Among them, that of Tom Beauchamp and James Childress is certainly the best-known and most important. While Beauchamp and Childress concede that there are influential theories of autonomous persons, they focus on autonomous choices from the outset, immediately marking a significant shift in focus (Beauchamp & Childress, 2019, p. 100). Their practical interest is directed toward deciding in individual cases—especially in the field of biomedicine—how actions are to be evaluated. In particular, they reject those approaches in which second-order preferences play a central role (focusing mainly on Gerald Dworkin's account of autonomy which is similar to Frankfurt's but was developed independently, cf. Dworkin, 1970). Instead of an "ideal theory," Beauchamp and Childress are concerned with autonomy under "nonideal conditions" (Beauchamp & Childress, 2019, p. 102). They explicitly start from the premise that "the everyday choices of generally competent persons are autonomous" (Beauchamp & Childress, 2019, pp. 100–102). While Frankfurt might agree with this, Kant apparently does not assume that we act predominantly autonomously. Very often, we seem to simply follow our inclinations. Strictly speaking, we can never be quite sure whether we really meet the high requirements of Kant's autonomy, since we are not fully transparent to ourselves. For Kant, it is sufficient that autonomous action is possible in principle. For Beauchamp and Childress, on the other hand, this is inadequate. For them, autonomous action is the default.

In their conception of autonomy, Beauchamp and Childress assume three constitutive elements, namely intentionality, understanding, and non-control (Beauchamp & Childress, 2019, p. 102). In this context, they understand intentionality as binary, whereas understanding and non-control allow gradations. Consequently, they also understand autonomy to be gradable, precisely as a function of the last two elements. In this, they differ clearly not only from Kant but also from Frankfurt.

In practical contexts, questions arise about, among other things, when a person has the competence to make an autonomous decision. Therefore, Beauchamp and Childress spend some effort discussing different standards. Eventually, they favor setting standards of incompetence (and using empirical tests for checking), which is in line with their basic assumption that people act autonomously most of the time (Beauchamp & Childress, 2019, pp. 115–118). They proceed similarly with regard to the question of how much information a person needs in order to be able to make an autonomous decision. In view of the principle of informed consent, which plays a central role in medicine and other fields, this is crucial for Beauchamp and Childress. Informed consent is precisely intended to ensure that individuals can make autonomous decisions. The authors distinguish the "professional practice standard," the "reasonable person standard," and the "subjective standard" (Beauchamp & Childress, 2019, pp. 123–125). Eventually, they conclude that "for purposes of ethics, it is best to use the reasonable person standard as the initial standard of disclosure and then supplement it by investigating the informational needs of particular patients or potential research subjects" (Beauchamp & Childress, 2019, p. 125). This is particularly interesting because it makes clear that Beauchamp and Childress assume that autonomy is something valuable in need of protection. They claim: "To respect autonomous agents is to acknowledge their right to hold views, to make choices, and to take actions based on their values and beliefs" (Beauchamp & Childress, 2019, p. 104). That these basic rights exist is beyond question for them. Most people today would certainly agree with this. Nevertheless, it is important to note that this view differs from that of Kant and Frankfurt. Kant, in particular, asks the more basic question of why certain choices or actions should be valuable and worth protecting. His answer is that they are only valuable if they are morally justified. If not, they are just natural processes which, taken by themselves, are neither valuable nor worth protecting. In other words: While Kant merges autonomy and morality and infers that autonomy is only valuable if it follows moral principles, Beauchamp and Childress assume that every human being has a right to self-determination which, in turn, must be protected. The only question that arises for them is whether a concrete decision was actually made knowingly and willingly. If that is the case it must be respected. Kant sees no reason for this, and Frankfurt mostly excludes the moral dimension in his approach.

In contemporary applied ethics, Beauchamp's and Childress's practical understanding of autonomy has been widely adopted. It is particularly suitable for establishing entitlements to protection and determining conditions for restrictions (e.g., in the case of limited capacity for understanding). Compared to Kant and Frankfurt, Beauchamp and

Childress start from a less demanding concept. For them, autonomy means de facto self-determination, whereby it is taken as self-evident that this is worthy of protection, which is in line with established fundamental rights. What all three conceptions have in common, however, is that they tie the concept of autonomy to the concept of person. They differ significantly in terms of when persons are autonomous and what that means exactly. But they agree that it is only persons who can be autonomous.

Luciano Floridi: The Power to Decide

Approaches that apply autonomy to artificial systems abandon this last unifying element. While this has been common in technical disciplines—admittedly often without realizing the profound change this means in terms of the traditional understanding of the concept—for some time, there are also conceptions from within philosophy which abandon the connection between autonomy and personhood. A prominent example of this more recent development is Luciano Floridi.

In his influential book *The Ethics of Information*, Floridi attempts, among other things, to arrive at a new understanding of the concept of an agent. First, Floridi introduces the notion of "level of abstraction (LoA)" to allow for different levels of analysis (2013, pp. 31–34). On one level of abstraction, according to Floridi, an agent is "a system, situated within and a part of an environment, which initiates a transformation, produces an effect, or exerts power on it over time" (2013, p. 140) According to this definition, earthquakes are agents, as Floridi points out. However, since this apparently is exceedingly broad, he suggests changing the LoA and including three criteria, namely "interactivity," "autonomy," and "adaptability" (2013, p. 140). For our purpose, the criteria "interactivity" and "adaptability" are less important. We, therefore, will focus on what Floridi means by "autonomous."

Floridi defines the term *autonomy* as follows: "*autonomy* means that the agent is able to change its state without direct response to interaction: it can perform internal transitions to change its state. So, an agent must have at least two states" (2013, p. 140). Floridi immediately supplements this rather parsimonious definition with the following explanation:

> Autonomy imbues an agent with some degree of complexity and independence from its environment and from those who built the agent. For example, the programmers of Deep Blue were only indirectly responsible for its win, since it 'learnt' by being exposed to volumes of games to such an extent that the programmers themselves were quite unable to explain, in any terms of chess parlance, how Deep Blue specifically played [. . .]. (2013, p. 140)[1]

From these remarks, it becomes clear how far Floridi's understanding of autonomy departs from the philosophical tradition. For him, two features are crucial for attributing autonomy: complexity and independence. Every natural or artificial system that has a certain (inner) complexity insofar as it can take on different states and is able to change

[1] Whether the chess computer Deep Blue, which won against the world champion at the time, Gary Kasparov, in 1997, is a well-chosen example is a question to be left open here. In any case, Deep Blue did not operate with today's machine learning methods (cf. Hsu, 2002).

them independently from external controlling interventions (i.e., can shift from one state to another), is autonomous according to his view. Floridi leaves no doubt that, in his understanding, many systems are autonomous that traditionally would not have been considered to have this property (2013, pp. 141–146). A pendulum, for example, is autonomous since it has two states between which it alternates without controlling influence. To be sure, a pendulum is not an agent since it is neither interactive nor adaptive. In contrast, (some) current computers are agents because they fulfill all three criteria, in particular being autonomous. Crucially, according to this understanding, autonomy has nothing to do with forms of self-determination—be it according to moral principles, as in Kant, be it according to self-chosen life plans, as in Beauchamp and Childress, be it according to higher-level volitions, as in Frankfurt. Moreover, autonomy does not indicate any kind of determinations of will but merely denotes degrees of complexity and independence.

Floridi goes on to introduce another subcategory, namely that of "moral agent" (2013, pp. 146–148). By this, he means agents capable of *morally qualifiable action*. Note that this does not imply that such agents are capable of taking responsibility. On the contrary, Floridi wants to escape the following dichotomy:

> - "moral agency, therefore responsibility, therefore prescriptive action, versus
> - if there is no responsibility then there is no moral agency, but without the latter there is no need for any prescriptive action" (Floridi, 2013, p. 159).

In other words, he worries that agents can perform actions that are morally relevant (i.e., can cause moral good or evil), but that no responsibility can be attributed. One can share this concern without accepting Floridi's notions of an agent and agency. As seen above, more traditional approaches have taken a different route: They have coupled the notion of an agent with a normative conception of autonomy. One might say, then, that the problem arises only because Floridi gives up this coupling. This is particularly relevant if the concept of autonomy is applied to artificial systems to which, at least so far, no one has been willing to grant the ability to assume responsibility. Any understanding of autonomy that simply refers to a form of independence runs the risk of making attributions of responsibility ambiguous. In the ongoing discussion about how to deal with AI, this problem plays a considerable role. Some years ago, Matthias (2004) already discussed whether the use of AI might lead to a *responsibility gap* (without, however, implying that AIs are autonomous agents). Whether such a gap really exists is still the subject of intense debate. If the concepts *agent* and *autonomy* are also applied to (non-responsible) artificial systems, then this at least threatens to create confusion. For authors like Beauchamp and Childress and even more so for all those who stand in the Kantian tradition, the terms autonomy and responsibility are inextricably linked.

Similarities and Differences

Onora O'Neill observed already 20 years ago:

> Contemporary accounts of autonomy have lost touch with their Kantian origins, in which the links between autonomy and respect for persons are well argued;

> most reduce autonomy to some form of individual independence and show little about its ethical importance. (2003b, p. 5)

There are now very different ideas of autonomy—both within and outside philosophy. While Kant advocates an extremely demanding notion of autonomy, which leads to the conclusion that only persons can be autonomous, but presumably only rarely are, Frankfurt's concept is more psychological in nature and largely omits ethical evaluation. Beauchamp and Childress, on the other hand, focus on ethical issues of everyday life. Their emphasis on intention, understanding, and non-control is evidently designed for practice. The key point is that they assume a right to self-determination which is self-evident. Finally, Floridi stands for a shift toward an understanding of autonomy that is guided by complexity and independence from control. In summary, we see here a conceptual pluralism according to which autonomy serves to characterize very different phenomena. In particular, autonomy can be meaningfully attributed to different entities in the world.

There is nothing wrong with such conceptual pluralism, at least as long as it does not lead to misunderstandings or erroneous conclusions. Moreover, there is no single academic discipline that can claim to have supremacy over concepts, and which is in a position to determine once and for all how a concept should be used. In the past, philosophy has occasionally claimed this privilege. However, today philosophy is no longer seen as the guardian of conceptual truth. Conceptual plurality is simply the result of, among other things, technological developments and shifts in societal attitudes. Still, philosophy is in a good position to highlight problems that may arise from conceptual pluralism. Especially in interdisciplinary debates, it is often philosophy that is most intimately involved with conceptual differentiations. Therefore, philosophy can help clear up conceptual mess, if necessary.

To be sure, conceptual clarification is not a purely descriptive enterprise. It is also an attempt to find out which conceptions are most appropriate. It is the task of the humanities and natural sciences to jointly evaluate different understandings of concepts. Importantly, when we talk about the autonomy of persons, our understanding is often still inextricably tied to moral considerations. Such considerations, though, are only meaningful against the background of specific understandings of autonomy. Therefore, we must be careful not to transfer normative implications to other usages of the word *autonomy*. Transfers like these can easily happen, when we assume that there is no difference between artificial agency and human agency or, likewise, between animal agency and human agency. They can also easily occur when we start to discuss the moral standing of AI, given that it is supposed to be autonomous and that autonomy is morally valuable.

The different concepts of autonomy relate to different assumptions of values. According to Kant, the value of autonomy is absolute, because autonomy and morality are simply two sides of the same thing. For Frankfurt, on the other hand, this is not the case. Autonomy merely denotes a kind of higher-level concordance of desires. For Beauchamp and Childress, however, autonomy also has a high value, simply because the individual right to self-determination is seen as a normative basis. The technical understanding, according to which autonomy merely denotes a kind of situational independence from external control, is neutral in evaluative terms. Autonomy understood in this way is completely decoupled from morals and values.

One can get the impression that in many current debates different dimensions of meaning of autonomy merge, sometime unconsciously and inexplicitly. When people talk about autonomous weapons systems, for example, they are referring to systems that operate temporarily without human control. If that were all, the discussion could possibly be limited to clarifying error probabilities and other technical details. However, the discussion about autonomous weapons systems is very intense, almost emotional. This may be related to the fact that it involves the potential killing of people, which is always a sensitive topic. It may also have to do with the fact that the attribution *autonomous* signals more. Meaning dimensions of responsibility and morality resonate—properties that artificial systems naturally do not have. Of course, this would have to be discussed in more detail to do justice to the complex subject. However, it is meant here only as a brief example of the fact that conceptual pluralism can lead to problems. Discussions about artificial intelligence and autonomous systems, in particular, would probably benefit from avoiding conceptual overtones.

One way to do this would be to avoid using the term autonomy altogether and replace it with more clearly defined terms. However, such approaches rarely prove successful. After all, the prohibition of using a term is all too reminiscent of a police measure that seems inappropriate in open discourse. It might be more fruitful to call for more conceptual transparency in discussions. Thus, at least in interdisciplinary debates, whenever autonomy or autonomous systems or autonomous decisions are mentioned, it should be disclosed what exactly is meant. This can sometimes be a tedious business. However, it can help to avoid misunderstanding and erroneous conclusions. Especially when it comes to ethical evaluations, we should not shy away from this additional effort.

Outlook

As in all kinds of pluralism, there is a benefit to conceptual pluralism. We should remain open to this benefit. At the same time, pluralism can lead to problems, and sometimes even to profound misunderstandings. In current debates, this is particularly evident with regard to the concept of autonomy. However, we should not conclude from this that the concept of autonomy must be abandoned. Rather, whenever we speak of autonomy, autonomous systems, autonomous decisions, or the like, we should try to make it as clear as possible what we mean (and imply) by this. And whenever someone else uses these expressions and we have the impression that we may have different meanings in mind, we should investigate what our counterpart exactly wants to say. In this way, conceptual pluralism can become a driver for interdisciplinary debate and, especially with regard to autonomy, may even help us to abandon outdated ideas. The accounts of autonomy by Kant, Frankfurt, Beauchamp and Childress, and Floridi are definitely not the only ones out there. Still, they offer an idea of how wide the conceptual spectrum is. They also provide a good starting point for further interdisciplinary exchange about autonomy, about what things in the world we call autonomous and what exactly we mean by that.

Author Biographies

Roman Wagner (PhD, University of Bonn) is a research assistant at the German Reference Centre for Ethics in the Life Sciences. His research connects foundational metaethical and metaphysical topics with questions of applied ethics.

https://orcid.org/0009-0009-3605-9876

Bert Heinrichs (PhD, University of Bonn) is full professor of ethics and applied ethics at the Institute of Science and Ethics (IWE) at the University of Bonn and Head of the Research Group "Neuroethics and Ethics of AI" at the Forschungszentrum Jülich. Before he served as Head of the Scientific Department of the German Reference Center for Ethics in the Life Sciences (DRZE), Bonn. In his research, he is primarily concerned with various topics in applied ethics. In addition, he also deals with questions of ethics and metaethics.

https://orcid.org/0000-0002-0181-0078

References

Allison, H. (2013). Autonomy in Kant and German Idealism. In O. Sensen (Ed.), *Kant on moral autonomy* (pp. 129–145). Cambridge University Press. https://doi.org/10.1017/CBO9780511792489.010

Beauchamp, T. L., & Childress, J. F. (2019). *Principles of biomedical ethics* (8th ed.). Oxford University Press.

Christman, J. (2014). *The inner citadel: Essays on individual autonomy*. Echo Point Books & Media.

Dworkin, G. (1970). Acting freely. *Noûs, 4*(4), 367–383. https://doi.org/10.2307/2214680

Engstrom, S. (2009). *The form of practical knowledge. A study of the categorical imperative*. Harvard University Press.

Floridi, L. (2013). *The ethics of information*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199641321.001.0001

Foot, P. (2003). *Natural goodness*. Clarendon Press.

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

Frankfurt, H. (1995). *The importance of what we care about. Philosophical essays*. Cambridge University Press. https://doi.org/10.1017/CBO9780511818172

Frankfurt, H. (2006). *Taking ourselves seriously & getting it right*. Stanford University Press.

Hsu, F.-H. (2002). *Behind deep blue. Building the computer that defeated the world chess champion*. Princeton University Press.

Kant, I. (1785). *Groundwork of the metaphysics of morals.* M. Gregor and J. Timmermann (Eds.). Cambridge University Press [page references according to the *Akademie-Ausgabe*, vol IV].

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*, 175–183. https://doi.org/10.1007/s10676-004-3422-1

O'Neill, O. (2003a). Autonomy. The emperor's new clothes. *Aristotelian Society, 77,* Supplementary, 1–21. https://doi.org/10.1111/1467-8349.00100

O'Neill, O. (2003b). Some limits of informed consent. *Journal of Medical Ethics, 29*, 4–7. https://doi.org/10.1136/jme.29.1.4

Oshana, M. (2006). *Personal autonomy in society*. Routledge. https://doi.org/10.4324/9781315247076

Pohlmann, R. (1971). Autonomie. In J. Ritter (Ed.), *Historisches Wörterbuch der Philosophie* (vol. 1, pp. 701–719). Schwabe Verlag. https://doi.org/10.24894/HWPh.367

Schneewind, J. (1997). *The invention of autonomy*. Cambridge University Press. https://doi.org/10.1017/CBO9780511818288

Sophocles. (1891). *The Antigone of Sophocles*. Edited with introduction and notes by R. Jebb. Cambridge University Press. http://www.perseus.tufts.edu/hopper/collection?collection=Perseus:collection:Greco-Roman