International Journal of Philosophical Studies

ISSN: 0967-2559 (Print) 1466-4542 (Online) Journal homepage: www.tandfonline.com/journals/riph20

Kant's Empowering Conception of Hybrid agency*

Carla Bagnoli

To cite this article: Carla Bagnoli (12 Sep 2025): Kant's Empowering Conception of Hybrid agency*, International Journal of Philosophical Studies, DOI: 10.1080/09672559.2025.2545760

To link to this article: https://doi.org/10.1080/09672559.2025.2545760

Published online: 12 Sep 2025.

INTERNATIONAL JOURNAL OF PHILOSOPHICAL STUDIES
https://doi.org/10.1080/09672559.2025.2545760

Routledge
Taylor & Francis Group

Kant's Empowering Conception of Hybrid agency*

Carla Bagnoli

Department of Education and Human Sciences, University of Modena and Reggio Emilia, Reggio Emilia, Italy

ABSTRACT

On Kant's view, human agents are 'animals endowed with reason,' sensitive to both natural and moral incentives. His model is hybrid rather than dualistic: insofar as humans are sensitive to the authority of rational norms, they can transform themselves from *animal rationabile* into *animal rationale*. This raises the question: how do hybrid agents navigate the heterogeneity of incentives? This is a live question, giving rise to a methodological puzzle. What methodology fits hybrid agency? Two opposing approaches, reductive naturalism and anti-naturalism, have prevailed, with John McDowell offering a non reductive variety of naturalism based on second nature. In contrast, this essay argues that Kant's view vindicates the positive aspects of naturalism and anti-naturalism, while rejecting the reductive ambitions of the former and the robust ontology of the latter. Hybrid rational agency, characterized by self-reflection, allows for rational changes that are not merely adaptive responses but radical, original acts of reorientation toward the world. By refocusing on Kant's account of moral agency, we recover the resources for a distinctive critique of reductive naturalism, showing its conservative and alienating effects. Kant provides an empowering view of human agency, in dialectical contrast to the second nature model based on habit and social training.

ARTICLE HISTORY Received 9 September 2024; Accepted 23 July 2025

KEYWORDS Kant; McDowell; naturalism; moral agency practical reason; respect

On Kant's view, human agents are 'animals endowed with reason' (A 7:321). Their animal embodiment makes them susceptible to natural incentives, interested in avoiding frustrations and in satisfying their desires and needs. Because they are endowed with reason, however, they are also capable of acting on principle, in accordance with the norms of rationality. This dual source of human agency is a recurring theme in philosophical discourse, but Kant proposes an important specificity that makes his own model a hybrid one, rather

CONTACT Carla Bagnoli carla.bagnoli@gmail.com

*In 2024, a previous draft of this paper was discussed at the 25th World Philosophy Congress *Philosophy across Boundaries*, held in Rome, and at the 14th International Kant Congress: *Kant's Project of Enlightenment*, held at the University of Bonn. I also benefitted from the workshop *Relations to Oneself and to Others*, held in May 2024 at Christ Church College, University of Oxford.

than dualistic: Insofar as humans are sensitive to the authority of rational norms, they can make themselves into rational agents, thereby transforming themselves from '*animal rationabile*' into '*animal rationale*' (A 7:321). I use the term 'hybrid' to capture the following problem in Kant's theory of rational agency: How do hybrid agents negotiate the heterogeneity of incentives?

This is a live question, and has given rise to a methodological puzzle: What kind of methodology fits hybrid agency? Two opposing approaches have prevailed in a debate whose terms are largely inherited from the Kant–Hume dispute (Skorupski 1993). On the one hand, naturalism recognizes no genuinely normative phenomena in nature. It proposes a reductivist account of moral action, generally in terms of a combination of cognitive and conative states: specifically, beliefs and desires. On the other hand, anti-naturalism holds that nature is infused with normativity, and that therefore there is no sharp ontological divide between matters of fact and matters of value. How to situate Kant's view within thi7s debate has remained an open question.

Current interpretations have not only revived this old dispute but also deepened and broadened its significance, in a way that makes the initial, blunt opposition between naturalism and anti-naturalism seem problematic and worthy of reconsideration. One prominent strategy aims to carve a middle way between reductive naturalism and anti-naturalism via the concept of second nature. John McDowell adopts this strategy to criticize and amend Kant's philosophy, which he takes to rest on a conception of nature wholly deprived of normativity. According to McDowell, this disenchanted view of nature misleadingly suggests that only anti-naturalism can fully vindicate the normativity of reason.

By contrast, I argue that we owe to Kant both the formulation of the philosophical problem of hybrid agency and a plausible solution to the methodological puzzle presented above: one which vindicates the positive cores of both naturalism and anti-naturalism by rejecting the reductivist ambitions of the former and the robust ontology associated with the latter. Hybrid rational agency is marked by self-reflection, which allows for rational changes that are not merely adaptive responses to the world, but also radical and original acts of reorientation toward it. By refocusing on Kant's account of moral agency, we can recover the resources for a distinctive and devastating critique of reductive naturalism that reveals its excessively conservative and alienating implications. Furthermore, we can show that Kant provides an empowering account of human agency, one which contrasts with the model of second nature based on habits and social training.

1. The Problem of Hybrid Agency

On Kant's view, human agents are hybrid ones: 'animals endowed with reason' (A 7:321), susceptible to natural incentives but also guided by, and sensitive to, the norms of rationality. This latter feature makes humans free, and for this very reason unpredictable, but also, and paradoxically, socially reliable – agents who can be trusted to act together. Unlike other hybrid models of agency, Kant's is marked by practical and inner freedom: Humans are free to change, bound by rational norms that guide them in exercising their freedom, and also capable of transforming themselves from animals merely capable of reason into rational animals.

To explain how these hybrid agents perform the required self-transformation, we need to articulate the relation between rationality and nature. Theories of hybrid agency come in two main varieties. The *additive* model represents the hybrid agent as layered, composed of different capacities or functions that operate separately and independently. Both animals and humans share the lower faculties, and therefore hybrid agency is what results from the addition of rational capacities. However, that addition does not alter the operations of the lower faculties or the scope of their achievements. This remains a fundamentally dualistic picture of the mind, with animality and rationality forever opposed to one another. By contrast, the *transformative* model represents natural and rational faculties as entangled, interacting at various levels of their operations so that reason actively shapes the exercise of the lower capacities.[1] Rather than simply augmenting the lower capacities, the rational and conceptual faculties transform them by enabling subjects to make judgments and evaluate objects in conceptual terms. Consequently, human desire and agency are not just more complex than their animal counterparts, but qualitatively distinct. On this account, human agency is hybrid not only because it involves faculties which both animals and disembodied beings lack but also because the various capacities of the embodied human mind cooperate so that they elevate beyond their individual roles through their collaborative activities. This cooperation produces a unique form of agency that is reducible to neither pure will nor animal instinct and cannot be understood as a combination of the two.

In my view, the transformative model provides the only genuine account of human agency's hybridity, as it highlights the various faculties' systematic interactions and interdependencies, and asserts that hybrid agency emerges from these faculties' collaborative and mutually influential activities within the embodied mind. In contrast, the additive view ultimately remains dualistic, as it posits that hybrid agency arises from a combination of distinct, independent faculties that remain separate throughout their operations. Hybrid agency, on the transformative account, is neither the mere sum of mental capacities and

embodiment nor the simple stratification of lower and higher faculties; it is a fundamentally distinct form of embodied rational agency.

2. Second-Nature Naturalism

The dichotomy between these additive and transformative models has gained prominence in a debate heavily influenced by John McDowell, whose aim is to find a middle ground between 'bald naturalism,' which denies any form of normativity in how our capacities operate, and 'rampant Platonism,' which consigns it to a supernatural realm (McDowell 1996, 1998). While bald naturalism seeks to 'domesticate conceptual capacities within nature conceived as the realm of law' (McDowell 1996, 73), McDowell aims to vindicate normativity without compromise, asserting that the 'contrast of logical spaces' is genuine and requires philosophical recognition (73). However, he also questions the divide between naturalism and anti-naturalism, diagnosing it as rooted in a misconception of nature. His remedial strategy is to introduce the concept of a 'second nature' (84): learned or acquired behaviors, skills, and habits that have become indistinguishable from one's original endowments.[2] On this model, 'first nature' refers to the potentialities of a normal human organism; their actualizations enable the human being to claim its 'second nature' (McDowell 1996, 84). The intricate relationships between these actualizations belong to the logical space of reasons (McDowell 1996, 19).

The claim that the animal and human hybrid is not an admixture affects the nature of the relevant norms. On the one hand, the norms of nature cannot be derived from truths about humans – contrary to the ambitions of reductivist naturalism. On the other hand, the norms of reason are not independent of and isolated from 'the merely human,' and do not 'float freely above the possibilities that belong to the normal human organism' (McDowell 1996, 109–10, and also 114, 118). This is how the concept of second nature carves out the space for 'relaxed naturalism' (89).[3] While many of McDowell's starting points are avowedly Kantian, he holds that Kant lacks 'a pregnant notion of second nature' because he locates spontaneity outside nature (McDowell 1996, 97)[4] via the transcendental constitution of empirical reality (McDowell 1996, 98), and thereby neglects the relationship between animality and humanity.[5] The concept of second nature is intended to explain the relationship between nature and spontaneity or normativity, addressing a tension that Kant leaves unresolved.[6] McDowell turns to Hegel for the view that nature itself falls within the conceptual sphere, and thus is inherently 'thinkable' (McDowell 1996, 28); he also favors the Aristotelian view that agents' conceptual capacities for excellence are both natural and educable McDowell (1979, 82). The 'resulting habits of thought and action,' built up by education, are second nature to human

beings, and therefore cannot 'float free of potentialities that belong to a normal human organism' (McDowell 1996, 84).

McDowell's argument largely depends on an interpretation of Kant's transcendental idealism that is disputable and has been disputed – in my view, successfully.[7] I concur with Allison (1997, 39) and others that McDowell's critique of Kant's transcendental idealism fails to appreciate its critical nature. In contrast to McDowell, I also share Allison's conception of the noumenon as a limiting concept 'drawing a transcendental (not an ontological) distinction between things as they appear (subject to the conditions of human cognition) and as they are in themselves (thought independently of these conditions)' (Allison 1997, 45–7). My aim is not to defend Kant's transcendental idealism nor to argue for its compatibility with a relaxed variety of naturalism.[8] Rather, I hope to contribute to moving forward by departing from this debate altogether, in order to refocus on Kant's account of moral agency.

3. Kant's Hybrid Account of Human Agency

While McDowell holds that Kant is driven away from naturalism by his ethical and religious concerns, I argue that his reflections on ethics and religion actually articulate a promising alternative to both naturalism and rampant Platonism.[9]

My task may seem daunting, if not hopeless. First, Kant's ethics is largely considered anti-naturalistic (Skorupski 1993), primarily for three reasons: because of his strenuous defense of the a priori method in ethics; because he claims that 'the ground of obligation here must not be sought in the nature of the human being or in the circumstances of the world in which he is placed' (G 4:389, 4:411); and because of his relentless attack on the view that morality is a matter of habits and customs, and thus conditional and hypothetical in its commands: '[the moral laws] hold as laws only insofar as they can be seen to have an a priori basis and to be necessary' (MS 6:215).

Second, Kant's view of human agency is largely regarded as dualistic rather than properly hybrid (Abela 2015, 56–73; Bader 2015; Louden 2020, 3–17; Timmermann 2022). Given the sharp divide separating animality from humanity and personality, it may be hard to see how Kant's view qualifies as transformative at all (McLear 2022, 112).[10] Third, and relatedly, his ethical theory is largely regarded as an exercise in ideal theory; the rational norms it posits govern embodied and disembodied agents alike, as though they were equally unbridled, free from the constraints of human embodiment (McDowell 1996, 114, 118).

Although these characterizations appear plausible, there are reasons to resist them. First, regarding the anti-naturalist characterization, Kant's conception of moral agency does not add anything supernatural in accounting

for rational agency, so to this extent, at least, his account does not violate any requirements of naturalism.[11] More importantly for the purposes of this essay, Kant's ethics qualifies as naturalist insofar as his account of the subjective authority of both the norms of reason and the moral law is related to human moral sensibility, which is inherent in and constitutive of humanity – even though their validity is neither contingent nor conditional on human nature. Moral sensibility explains both why embodied rational agents face normative questions and how they are bound by the norms of reason in responding to them.[12]

Second, the seemingly powerful argument for the dualistic view involves the very separation between the dispositions of animality and humanity. The view's adherents note that 'animality' refers to the lower or sensible faculties and to the instinctual drives, such as the desire for self-preservation, reproduction, and community; these are associated with a mechanical variety of self-love (R 6:26–7), for which reason is not required. However, the real question concerns whether brutes and animals endowed with reason experience the same kind of instinctual self-love. In contrast to the suggestion that they do, it is arguable that for human animals, self-preservation, reproduction, protecting one's offspring, and a longing for community are shaped by distinctly normative practices that require normative membership of such a community. These practices, often highly ritualized, are not simply a more complex version of how non-human animals satisfy their instincts; rather, they transform these instinctual drives into activities that can be evaluated as right or wrong, according to normative standards. The analogy with non-rational animals' instinctual drives is weak and holds only at a functional level. Humans' natural drives, unlike brutes', are governed by (rational, social, and moral) norms, can be either expressed or inhibited, and are culturally shaped in ways that make them look very different from merely 'physical' or 'mechanical' feelings. They differ from brutes' in kind, and this supports the transformative interpretation.

The mere separation of faculties can offer only a partial argument against the transformative view. Hybrid agents are defined by the multiple and separate faculties they possess; the outstanding issue is whether these faculties' operations are all mutually independent, or whether, at least when executing some operations, they cooperate so tightly that each one's contribution becomes indiscernible from the others'. The working assumption behind the argument from separation concerns the respective determinations' oppositional nature: if they are conflictual, it is unlikely that they could be working together and in ways that overcome their separation. Kant's texts seem to offer ample evidence of this conflict: 'we do not achieve the perfection of humanity in the determination of animality, and if we want to achieve the perfection of humanity, then we must do violence to the determination of animality.'[13] Such references do point unequivocally to the heterogeneity of

incentives. However, it remains disputable that they decisively support a dualistic model of agency, with evil originating in the direct opposition between the dispositions of humanity and animality. On the contrary, there are both textual and philosophical grounds supporting the hybrid interpretation.

In the next section, I argue that Kant's account of moral agency's emergence out of the normative dynamics of incentives brings into sharp relief how Kant's hybrid model centers on the transformative power of rationality.

4. The Mark of Moral Agency

The hybrid view holds that rationality affects the ways in which animality is actualized – through incentives that can be the content of the subjective maxims of action – and that animality, in turn, affects the way in which rationality is actualized. However, natural incentives do not directly determine motivation: unlike brutes, hybrid agents can employ reason to decide between conflicting incentives. In the opening section of *Religion*, Kant rejects the 'combat model,' according to which incentives oppose each other directly; he thus adopts a hybrid model of rational agency that makes evil a matter of choice, and therefore imputable. A key aspect of this theory is the so-called 'incorporation thesis,' which denies that desires, as such, suffice to explain actions. For a rational agent, desires are merely possible grounds for making choices (see Wood 1999, 51–3). I take this to be the anti-dualistic core of the proposal. While the incorporation thesis is explicitly formulated in *Religion*, it also 'underlies virtually everything that Kant has to say about rational agency' (Allison 1990, 40).[14] The incorporation thesis helps us understand why evil is imputable even though it is radical and inextirpable. Kant accounts for the imputability of evil by presenting it as an operation of the will, the faculty that negotiates heterogeneous incentives: evil consists in the misarrangement of incentives, and this results from the activity of the will. The predisposition to rank the natural incentives above the moral incentive is universal and individuated: 'universal in the species, but original for each individual [*allgemein in der Art, aber ursprünglich bei jedem Einzelnen*]' (R 6:31). Every member of the human species is subject to the same radical tendency, but evil inheres in each person's own free choice, not in some collective natural drive.

The possibility of choosing evil rests on a peculiarity of the human condition. While the same normative standards of reason apply to all rational beings, regardless of their embodiment, they apply to hybrid agents in a distinct way, precisely because such agents are both finite and embodied.[15] Embodiment presents a significant challenge to self-knowledge. Humans lack full, direct cognitive access to their own mental states – such as intentions, beliefs, and dispositions toward action. As a result, self-observation or

introspection can never provide any complete insight into the self. Furthermore, self-love can obstruct self-understanding (G 4:407; R 6:51/71, 6:71/87–8; MS 6:447). This dynamic gives rise to self-deception – 'the inner lie,' fundamentally a moral pathology (MS 6:430–1). In the case of hybrid agency, the standard of the moral law is *felt* to be authoritative: The moral incentive is a subjective feeling, and feelings are experienced in those pleasurable and painful fashions which are available to embodied agents.[16] Unlike pathological feelings, the moral feeling arises when reason elicits it, not when bodily sensations prompt it. This claim does not rest on any supernatural ontology; it merely establishes that a feeling whose intentional object is formal can be elicited through reflection, unlike feelings that are aroused by material and external objects. Importantly, the specificity of human embodiment is involved in the operations of the moral feeling: one experiences that feeling as pleasant (e.g. the thrilling, self-comforting, and empowering feeling of overcoming one's inclinations), or painful (e.g. the shameful feeling of succumbing to inclinations, the humiliation of self-conceit, or the frustration of one's desires), or a combination of the two. No mysterious metaphysics is needed to account for these effects – nothing beyond one's recognition of the normative powers associated with self-reflection. Such powers are not only reflexive, in that they are exerted upon the very subject who exercises them, but also transformative, in that reflection changes its subjects, allowing them to claim responsibility for action.

Therefore, Kant's hybrid view of rational agency centers on the claim that exercising rational capacities allows for a variety of agency that, in self-reflection, is activated by a distinctive kind of representation of oneself *as free*.[17] Kant's account of reason's activity is directly related to his characterization of a human being as a practical subject. The distinctive mark of any rational activity is that one can engage in it only under the representation of oneself as free. Every rational being has a will and can act only under the idea of freedom. While freedom is a mere idea, it is regulative and thus, in its practical sense, a necessity: rational beings can act only under this idea (G 4:448). Thus, adopting the stance of agency is not itself a matter of choice, but necessarily related to representing oneself as free, both from external causes and from initiating action spontaneously.[18] Consequently, animals endowed with reason are not merely more capable of exercising their will or able to access more sophisticated rational tools. Rather, they qualify as a distinct category of agency, differing both from other animals and from disembodied rational beings or a 'holy will' (G 4:414, 4:439).

Fundamentally, rationality enables animals endowed with reason to represent themselves under the idea of reason. Via this self-representation, rational beings know themselves to be capable of genuine agency – capable of authorizing action, and thus morally imputable for what they do. This self-representation is not solely an act of intellectual cognition; it is felt under the

guise of the moral feeling of respect. This feeling, in its most abstract application, should thus be regarded as the mark of moral agency. Like natural feelings, it is a subjective condition of the mind. Unlike natural feelings, however, the moral feeling is originated by a pure act of will, rather than in response to an external object because of, say, its desirability.

The moral feeling of respect is, in one of its fundamental aspects, a self-directed feeling: the subjective experience of one's own freedom, understood in a practical, not a metaphysical sense – as autonomy. It does not merely accompany the idea that one acts freely; it marks the subjective experience of acting freely, undetermined by external factors. This feeling is necessarily associated with one's conception of oneself as a free agent; it is generated by the representation of oneself as capable of initiating an action out of nothing – both prior to and independently of natural drives, rather than in an attempt to satisfy pre-existing desires. Such a feeling is independent of perceiving something external to oneself as an object of choice; it is directed toward one's capacity for self-determination rather than any external target. However, it is not restricted to oneself; the feeling of one's own autonomy and self-determination is relational, and its complementary aspect is respect for others as likewise capable of self-determination. Thus, while it is a subjective feeling associated with one's own internal experiences, it is also, and at the same time, the feeling that somebody else is real and enjoys the same freedom as oneself.[19]

In its capacity as 'the sole moral incentive' (CPrR 5:78), respect enables subjects to act on reason, thereby establishing their authorship of what they are doing and their authority over it: this is the distinctive mark of moral agency the ground of moral imputability and reciprocal accountability. It is therefore in the exercise of moral agency that the hybrid character of human agency best emerges, as the actualization of animality, humanity, and personality altogether. Moral action best expresses what humans are capable of, insofar as they are animals endowed with the capacity for reason (A 7:321). More fundamentally, the moral problem of what one ought to do arises only for hybrid agents such as humans. It does not arise for animals, insofar as they are conceived as fully determined by natural laws – as beings whose acts can be fully explained by the mechanics of natural incentives. Nor does it arise for disembodied rational beings, whose wills are fully determined by the laws of reason and thus require no command to act on duty. It is *only* for human agents, insofar as they are hybrid, that the laws of reason are normative – that is, regulative rather than determining. Insofar as they are hybrid, human beings stand in contrast both to 'a holy will,' which is not 'capable of any maxim conflicting with the moral law,' and to animals that lack the rational capacity to formulate maxims.[20] For hybrid agents, such as humans, 'the relation of such a will to this law is dependence under the name of obligation, which signifies a necessitation, though only by reason and its

objective law, to an action which is called duty' (CPrR 5:32). The moral law applies to all rational beings as such, but embodiment affects how it applies.[21] The moral law governs a hybrid agent's mind and actions through imperatives which 'are only formulae expressing the relation of objective laws of volition in general to the subjective imperfection of the will of this or that rational being, for example, of the human will' (G 4:414).

To wit, humans are uniquely susceptible to the moral feeling of respect because of their hybrid nature. This feeling is the subjective way in which hybrid agents experience the authority of the norms of reason. This is an experience that sets them apart both from other animals and from disembodied rational beings, who can be fully and immediately determined by the laws of reason.

5. Moral Agency and Agential Authority

The argument supporting the hybrid and transformative interpretation conception is meant to rule out the objection that Kant's view of moral agency lapses into anti-naturalism. However, it does not domesticate or trivialize Kant's commitment to moral obligations' a priori, unconditional nature. On the contrary, it highlights why his stance rules out reductive naturalism – the view that the norms of practical reason are derived or abstracted from human nature, considered both as a biological and as an ontological entity. In sharp contrast to this view, Kant holds that 'moral concepts [. . .] cannot be abstracted from any empirical and therefore merely contingent cognitions' (G 4:411). My argument in Sections 2–4 shows that the sort of considerations that rule out the reductive variety of naturalism are best appreciated by moving away from the metaphysical dispute I discussed at the outset.

The reason is that the naturalistic derivation does not stem from positing a distinctive, irreducible moral ontology nor from claiming that norms of rationality float free of every natural anchoring. Instead, the problem is that treating rational norms as derivative from human nature undermines reason's autonomy and nullifies its self-authenticating authority.[22] Such a treatment also fosters a conservative and defeatist view of human agency, as it denies the possibility of initiating a moral transformation that disrupts existing conditions instead of merely responding to them. Rejecting naturalism on these grounds does not imply an additive or metaphysical conception of rational agency. Rather, it requires that such agency be (subjectively) represented as free from external determination.

For any action to be autonomous, this kind of *practical knowledge* must be in place.[23] For moral deliberation to be successful, it is not sufficient that the principle underlying one's action is prescriptive; it must also be self-legislated, and action-guiding insofar as agents take it to be so. This is what

ensures the autonomy of agents in moral deliberation; the prescriptive or conative feature does not distinguish autonomous actions from submissive or heteronomous ones and does not elucidate the relationship between one's knowledge of oneself as a practical subject and one's knowledge of one's duties.[24]

Subjects capable of rational choice stand in different relations to their own actions, and such relations can be made explicit by identifying their relevant underlying principles. But how? One might proceed from the inside out, searching for the inner determinants of virtue; or from the outside in, taking actions to be the criteria for any ascription of principles. Kant considers whether agents may acquire self-knowledge inferentially, looking backward at the patterns of their actions (R 6:71, 77); and given his appreciation of epistemic limitations, one should expect that the theoretical role such principles play will be modest.[25] Since a multiplicity of principles may serve as the basis for action, the inferential pathway from actions to principles is both speculative and problematic. Nonetheless, such an exercise is not entirely useless, provided that its aim is not to track the principles of action outside-in. Indeed, understood as a normative practice of self-scrutiny, going back to the maxims of action may be quite important – provided that this exercise is intended to regulate one's conduct, not to achieve evidential knowledge of one's own intentions.[26] This is how humans typically perform a sort of practical self-surveillance, which differs substantially from self-monitoring practices because it aims at claiming responsibility for action, not at attributing veridical states of mind to oneself.

All forms of self-knowledge are principled activities, but principles play different roles depending on whether these activities are deliberative or reconstructive, allowing practical subjects to be both agents and spectators of their own actions. In the former case, practical knowledge is *effective* in action; in the latter, it is knowledge about action or the mechanisms of agency. While acting, rational agents are guided by their practical knowledge of themselves as self-determining agents, thereby acknowledging and enacting principles' practical significance. This practical knowledge is related to their capacity to authorize action or *claim* moral responsibility for it – in contrast to a sort of speculative knowledge of the mechanisms that explain how an action occurs as an event occasioned by an individual's character traits and dispositions. Such knowledge can play a role in guiding action because it is directly related to the subjective maxims of action. Maxims encapsulate the subjective grounds of action, and thus carry a distinctively first-personal claim to moral responsibility. This is why the first-personal and action-guiding nature of self-knowledge is key to explaining autonomous action, and therefore the authorial dimension of moral striving.

Claiming moral responsibility for action in the first-personal mode does not entail epistemic transparency; the kind of knowledge at play in first-

personal accounts of intentional action and moral striving is not evidential. Accordingly, the privileges of the first-person stance do not concern the epistemic accessibility of one's own states, but one's agential authority to make legitimate claims and acknowledge responsibility. By contrast, when judging oneself as the agent behind a past action, one takes responsibility for that action by attributing responsibility to oneself in the third-personal mode. This activity appeals to the same principled contents of the subjective maxims of action, but these work as explanatory principles, evinced by an action's performative aspects: they can tell us nothing more about the moral fabric of the action than the agential patterns as they appear to an external observer.

One might call the latter kind of self-knowledge practical even though it is inert, on the grounds that it carries knowledge of the outward, patterned dimension of action. Presumably, patterns reveal actions' outward structure, including their consequences, although not the agent's intentions. In this case, the term 'practical' is used in a deflationary sense, to say that it concerns action, not that it guides action directly and overtly. However, even in this sense practical knowledge may not be completely ineffectual, and it may play a practical role, albeit indirectly: The prospect that a future authority might judge one's entire life may exert a certain normative pressure, prompting individuals to align their actions with their professed principles. '[S]ince he can derive no certain and definite concept of his disposition through immediate consciousness but only from the conduct he has actually led in life, he shall not be able to think of any other condition of being delivered to the verdict of a future judge [. . .] than that his whole life be one day placed before the judge's eyes, and not just a segment of it, perhaps the last and to him still the most advantageous' (R 6:77). In this case, the judge's stance is not observational but practical. Yet this only shows that the ethical ideal may have a grip on human character also through the representation of an authoritative moral exemplar.

These tight connections between rational agency, agential authority, and moral responsibility constitute the background against which the reasons for rejecting naturalism are best appreciated. To begin with, the very attempt to search for an exhaustive explanation of agency oversteps the proper bounds of reason, as it involves embarking on an investigation that aims to make human action determined and predictable via naturalistic patterns. This ambition runs against the critical project of a credible and 'scientific' metaphysics. The previous sections have also brought fully into the light a different reason to reject naturalism: seeking a complete explication of human agency amounts to knowing the determining causal grounds of action, and thereby to denying human agency its freedom. Therefore, a reductivist naturalist model not only overestimates reason's cognitive and predictive powers but also implicitly denies its free exercise.[27]

Unlike both reductive naturalism and the additive varieties of non-reductive naturalism, on Kant's account the exercise of rational agency does not merely result in a patterned performance nor is it reducible to observing rules or commands. It requires the agent to adopt a maxim in the first-person stance, thereby claiming moral responsibility for it. When agency is marked by the moral feeling of respect, it carries a distinctive kind of knowledge of action, and of oneself as an agent. While this kind of knowledge is not based on evidence and refers to nothing outside the subject, it represents a distinctive, and crucial, cognitive and practical activity. It is distinctive because it is the result of engaging in action in the first-person stance. It is crucial in that it is knowledge of the relation of authority between the agent and the principles underlying action; it therefore accounts for the paradigmatic way in which agents claim and take responsibility for action.[28] While it is possible to adopt other stances to explain how one acts, such an exercise cannot capture the distinctive relation of authority that agents establish with their own actions. For a rational being, conceiving of oneself as a free agent is a matter of necessity; to regard oneself solely as determined by natural propensities or external social structures therefore carries self-alienating effects.

Based on these considerations, it is arguable that reductive naturalism, which aims to explain actions as events in a natural causal structure, both clashes with practical subjects' representation of themselves as the genuine agents of their actions and fails to capture the relation of authority that hybrid rational agents establish with their actions under the idea of freedom.[29] This failure alienates agents from their deeds, inhibits their claiming responsibility for what they do, and thereby transforms others' normative claims and expectations regarding the agent's action into external, burdensome, and alienating constraints. Naturalism severs the three links that keep an ethical community together: those between rational agency and autonomy, between autonomy and moral responsibility, and between moral responsibility and social accountability.

6. The Possibility of Change and the Duty of Self-Transformation

Reason is a capacity whose exercise is a duty. Both undermining its independent authority and restricting the conditions of its free exercise are crimes against humanity (E 8:40). The very possibility of an ethical community rests upon the free exercise of reason, as does moral progress. Both of these collective ethical tasks depend on individuals' capacity for moral improvement and rational choice. This emphasis on self-transformation as a moral duty clashes with the conservative varieties of naturalistic determinism.

However, it also appears to diverge from second-nature naturalism, which not only underscores the educability of the conceptual capacities that allow

for excellence in action but also credits habituation as the fundamental source of normativity: '[The] ethical is a domain of rational requirements, which are there in any case, whether or not we are responsive to them. When a decent upbringing initiates us into the relevant way of thinking, our eyes are opened to the very existence of this tract of the space of reasons. Thereafter our appreciation of its detailed layout is indefinitely subject to refinement, in reflective scrutiny of our ethical thinking' (McDowell 1996, 82).

For Kant, moral transformation is of the mind, not of one's habits, and is fundamentally 'the firm resolve' to comply with duty (R 6:46). Formulaic practices, habits, and traditions are often depicted as hindrances to autonomy, even though they are modes of growing out of one's animality: 'It is difficult for any individual to work himself out of the immaturity that has almost become second nature to him. [...] Statutes and formulae, those mechanical tools of a rational use, or rather misuse, of his natural endowments, are the shackles of a perpetual state of immaturity' (E 8:36). Nonetheless, Kant not only acknowledges habituation's positive role in moral education but also provides an account of its role in moral transformation-one which is unique in the following respects. First, Kant identifies habituation's function by distinguishing two explanatory standpoints in moral change. In the abstract, the relevant kind of moral change is a conversion: a self-transformation that restores the priority of the moral disposition. Logically, the order of explanation for such a moral transformation is from mind to character to action; but as a matter of fact, for embodied and socially embedded rational agents it may happen in the reverse order, as a slow reform of one's habits (R 6:48). Discipline that instills good habits can lead to such a radical change of mind, and often does. These remarks on habits in moral education are justified by the limitations and the opportunities which human embodiment and social embeddedness create.[30]

Second, Kant identifies the positive effect of habituation in terms of the agent's observance of the law, which can be obtained 'through a gradual reformation of conduct and consolidation of his maxims' (R 6:47). This 'legal' sense of virtue is attainable without the agent's undergoing 'the slightest change of heart'; and since intentions are opaque, one cannot infer a moral conversion because one has observed a change in an agent's practices. Even so, legal virtue marks a significant achievement because, especially in the case of duties of right, it warrants compliance without resort to external incentives, such as force, deterrence, or sanction. Duties of right correspond to 'what we owe to each other'; unlike duties of virtue, they can be legitimately and coercively enforced from the outside. Legal virtue ensures that they are enforced by habituation rather than by deterrents or by the application of sanctions. When this effect is brought to the fore, one can appreciate how important habituation is for establishing relations of mutual

respect and recognition. To this extent, habituation's function in hybrid agency is akin to McDowell's (1979, 1995) second nature.

This account of habituation has three merits. First, it coheres with the acknowledgement that moral perfection is a regulative ideal for embodied rational agents: even though moral improvement is a duty, we always begin from the beginning.[31] Second, it emphasizes that a firm commitment to adopt the moral ideal is a personal moral responsibility that each rational agent should claim, without denying that such a path is communal and typically involves social practices of reasoning and education. Third, it recognizes the importance of habituation, while underscoring that a change in habits does not guarantee the kind of stability that is required in order to establish an ethical community. Legal virtue allows for practices of mutual accountability that enforce a system of reciprocal respect for individual freedoms. In contrast, moral virtue ensures the proper kind of stability, which does not depend on any external enforcement (whether through habituation or through deterrents), but on recognition of the equal normative status of persons as moral claimants who are worthy of respect.

Notes

1. (Boyle 2016), 4 (Conant 2016; McDowell 2009),; 272 (Pendlebury 2021).; McLear (2022) finds the term 'transformative' ambiguous and proposes a distinction between 'essentialist' and 'actualist' transformative views of rationality. According to the former, possessing rational capacities changes the essence or nature of an animal's other capacities; according to the latter, possessing rational capacities changes other capacities' conditions of actualization, or the content of their acts, or both. I use the term 'transformative' in the latter sense.
2. McDowell distinguishes other two senses of nature: the disenchanted nature of modern natural science (1996, 70–1) and the concept of nature as opposed to what lies outside it – the supernatural (cf. 1996, 77–8). Both senses are present in Kant's work.
3. This is also named the 'naturalism of second nature' and 'naturalized platonism' (McDowell 1996, 91).
4. See the third antinomy of pure reason, arising from the theses that (a) it is necessary to assume a causality of freedom, an absolute spontaneity of causes within nature (CPR B472, B474), in order to explain human thought (B 574–5) and action (B 476); and (b) 'everything in the world happens according to the laws of nature' (B 473), and thus there appears to be no freedom.
5. 'Greenberg and Willaschek quote Kant as saying "everything in the world happens according to the laws of nature". Reformulated in the terms I have been using, this would be the claim that every occurrence – everything that happens – can be described in terms that enable it to be subsumed under law. Formally speaking, this thesis leaves room for the move I have identified with placing spontaneity in nature; we could say that some of those occurrences – all of which, under other descriptions, are subsumable under law, according to the thesis – are also characterizable in the contrasting terms, terms that imply

responsiveness to reasons and resist subsumption under law. This would be a position on the lines of the domestication of Kant proposed by Davidson in "Mental Events" (1980b). But how genuine a space for spontaneity is provided by this formal move?' (McDowell 1996, 102).

6. '[. . . A] tension between nature and freedom arises as a potential problem for me because the conception of experience that I urge, in response to the dilemma I begin with, requires that capacities of spontaneity – freedom – can be actualized in operations of sensibility, which is surely part of our natural endowment. [. . .] I aim to make room for the required combination of freedom and naturalness by rejecting the equation between what is natural and what can be made intelligible in terms of its conformity to natural law' (McDowell 1996, 100).
7. Against the two-world construal of Kant's transcendental idealism, see (Allison 1990, 1997; Bird 1998). 'It is not that there really are two distinct realms, a natural, empirical realm and a non-natural, transcendent realm of things-in-themselves; rather there is just one realm, empirical reality, and a conceptual apparatus which tempts us to transcend it' (Bird 1995). Bird argues that 'McDowell not only draws no distinction between the transcendental and the transcendent, but also conflates them' (Bird 199, 228). Kant's claims are not 'empirical, causal, or psychological, but modal and epistemic' (229–32).
8. On the latter issue, see (Bagnoli 2022). In support of Kant's transformative view in the philosophy of mind, see (Conant 2016); (Pendlebury 2021), and (Boyle 2016, 2024).
9. At least, against the varieties of naturalism that aim at completeness in their accounts of human agency. But contrary to McDowell's assumption, there may be fairly reductivist accounts of the mind that can make sense of hybrid agency. The naturalistic model of hybrid agency D'Amasio (2021) defends does not purport to provide a complete account of agency – and in this sense, it is not reductivist – but it certainly proposes to make all the explanations it does include completely naturalistic. Reductivism takes explanation to be a thoroughly naturalistic and reductivist endeavor, though it may not presume that it can explicate all there is; some portion of reality may be inexplicable.
10. 'One would think that if he endorsed ETR [Essentialist Transformative Rationality], human animality would not be a central threat to our rational action, because it would itself be informed – or rather transformed – by our rationality. Instead, the most plausible reading of the texts is that our animality is in opposition to our rational nature, not an expression of it' (McLear 2022, 112).
11. The formulation and resolution of the third antinomy both reject the idea that moral agency can, or should, be explained through supernatural causation, as this would undermine the law-governed order of nature. The discussion of miracles in *Religion* aligns with this argument, so there is no ambiguity in the underlying methodological framework.
12. Kant's claim that practical reason is unconditional does not have to rest on any queer metaphysics; it can be explained by reference to reason's universal authority and recognizability (see O'Neill 2004). Based on 'the common idea of duty and of moral laws,' 'everyone must grant' that a moral law 'must carry with it absolute necessity' (G 4:389).
13. *Anthropologie Friedländer* 25:682. McLear interprets this as locating the origin of evil 'in the opposition between humanity and animality, or between the

physical, natural predispositions and the moral ones; the inevitable evil in the determination of the human being is the spur toward the good that the human being must perform' (McLear 2022), 112–3). McLear also cites *Anthropologie Pillau* 25:736 (1777–8), *Anthropologie Mrongovious* 25:1420 (1784–5) and 25:1342–3, *Religion* (R 6:32, 36/1796), and *Conjecture* (CBHH 8:115). I am less convinced that Kant holds the oppositional view in *Religion*; although the emphatic rhetoric of some passages in Chapter 1 may seem ambivalent, the subsequent chapters align with the transformative view.

14. In my view, the incorporation thesis coheres with and supports a constructivist conception of practical reasoning (see Bagnoli 2021). This contrasts with how some other constructivist interpreters take it to be a form of 'reflective endorsement': 'the reflective mind must endorse the desire before it can act on it, it must say to itself that the desire is a reason. As Kant puts it, we must make it our maxim to act on the desire' (Korsgaard 1996, 19). There are dissenting voices: Frierson questions the scope of the incorporation thesis and restricts it to desires stemming from the higher faculty of desire (Frierson 201, 63); McCarty argues against this interpretation of *Religion* because it underplays Kant's view that free choice belongs within the noumenal world (McCarty 2008, 247–8).
15. Kant often insists on the peculiarity of human agents in this regard: the concept of taking an interest in action presupposes a limitation in a being's nature, as do the concepts of maxims and incentives, in that the subjective constitution of the being's choice does not, of itself, accord with the objective law of practical reason. These concepts do not apply to the divine will (CPrR 5:79, 4:412–3n, 460).
16. Moral feeling is the capacity to take an interest in the moral law (CPrR 5:80). It makes us 'aware of the constraint present in the concepts of duty' (MM 6:399; see Bagnoli 2021).
17. The distinguishing feature of human nature is freedom, understood as a 'transcendental predicate' of a specific kind of causality (CPrR 5:94). However, this kind of freedom pertains to human self-representation: it is a rational necessitation, meaning one cannot but think that one's will has it (G 4:455). This is related to our self-consciousness of our susceptibility to the moral law.
18. To say that rational beings act under the representation of freedom is to say that they cannot but conceive of themselves as free *agents*; see footnote 16, about taking interest in action, in contrast to acting for interest. Their adoption of the stance of agency is a matter of rational necessity, unlike agents adopting the intentional stance in naturalistic accounts of agency, such as Dennett's (1987), 41). Against similar naturalistic interpretations of Kant that categorize his views as compatibilist, see (Allison 1990), 76–82.
19. This statement requires an argument; see Bagnoli (2021).
20. 'The very concept of duty is already the concept of a necessitation (constraint) of free choice through the law. [. . .] Such constraint, therefore, does not apply to rational beings as such (there could be holy ones) but rather to human beings, rational natural beings, who are unholy enough that pleasure can induce them to break the moral law, even though they recognize its authority; and even when they do obey the law, they do it reluctantly (in the face of opposition from their inclinations), and it is in this that such constraint properly consists' (MS 6:379; cf. MM 6:399).

21. The moral law 'does not hold only for human beings, as if other rational beings did not have to heed it' (G 4:389).
22. On the centrality of the question what a human being is, see Logik 9:25, 'Letter to Stäudlin,' May 4, 1793.
23. Pure practical reason constructs its proper objects through the moral law, which is the causality of freedom (5:49). My interpretation situates Kant's ethics within the dialogue with Aristotelian practical cognitivism, but it also facilitates a constructivist interpretation of the latter; see Bagnoli 2013.
24. O'Neill's emphasizes that maxims are 'prescriptive,' but this is not sufficient to explain the distinctively authorial dimension of moral striving which she aims to capture: 'Seen from the practical standpoint, maxims identified as the prescriptions agents strive to follow, whose enactment in varying circumstances is the object of their practical deliberation and their moral striving. [. . .] Virtuous action does not require us to know which principles we have internalized, but only which principles we are striving to live up to' (O'Neill 1996, 95). In my view, it is instead the distinction between the first-personal and third-personal modes of self-knowledge that explains the contrast that matters to O'Neill.
25. 'More generally we must distinguish the significant contribution maxims can make to practice – to guiding action – from their meager contribution to theoretical explanation' (O'Neill 1996, 90).
26. The duty of self-knowledge is a duty of self-scrutiny: the duty to be reflective, and hence to undergo the self-authenticating process guided by reason (O'Neill 1989, 173).
27. Inner dispositions and propensities are inexplicable (R 6:57, 59).
28. They cannot but conceive of themselves as free agents (G 4:448); cf (Stroud 1994).
29. The experience of the moral feeling and the presence of the moral incentive of respect show the moral law to be 'valid for us as human beings, since it arose from our will as intelligence and so from our proper self' (G 4:461).
30. A proper appreciation of social embeddedness allows us to recognize that Kant's conception of autonomy neither reduces to self-sufficiency nor excludes the role others play in the process of learning to become autonomous, see (Bagnoli 2020), and cf (Thompson 2004). However, I argue that learning through active engagement in practical reasoning as a social practice is more significant than mere habituation.
31. 'For what in our earthly life (and perhaps even in all future times and in all worlds) is always only in mere becoming (namely, our being a human being well-pleasing to God) is imputed to us as if we already possessed it here in full. And to this we indeed have no rightful claim (according to the empirical cognition we have of ourselves), so far as we know ourselves (estimate our disposition not directly but only according to our deeds), so that the accuser within us would still be more likely to render a verdict of guilty' (R 6:75–6; cf. MM 6:409, 405, 397).

Acknowledgement

I would like to thank Jeremy Fix, Christopher Horn, Evgenia Mylonaki, Frederick Neuhouser, Thomas Pendlebury, Oliver Sensen, Jens Timmermann, and the journal referees for their insightful comments on a previous draft.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the University of Modena and Reggio Emilia's Fund for Interdisciplinary Research FAR 2023 (Project: Conflict Resolution by Ethical and Religious Authority – CUP E83C23002460001).

ORCID

Carla Bagnoli http://orcid.org/0000-0003-2795-3795

References

Abela, P. 2015. "Kant, Naturalism, and the Reach of Practical Reason." In *The Transcendental Turn*, edited by S. Gardner and M. Grist, 56–73. Oxford: Oxford University Press.

Allison, H. 1990. *Kant's Theory of Freedom*. Cambridge and New York: Cambridge University Press.

Allison, H. 1997. "'We Can Act Only Under the Idea of Freedom'." *Proceedings and Addresses of the American Philosophical Association*, 71, 39–50.

Bader, R. M. 2015. "Kantian Axiology and the Dualism of Practical Reason." In *The Oxford Handbook of Value Theory*, edited by I. Hirose and J. Olson, 175–202. Oxford: Oxford University Press.

Bagnoli, C. 2020. "Equal Standing and Proper Reliance on Others." *Theoria* 86 (6): 821–425.

Bagnoli, C. 2021. "Respect and the Dynamics of Finitude." In *Respect*, edited by R. Dean and O. Sensen, 121–139. Oxford: Oxford University Press.

Bagnoli, C. 2022. "Kantian Constructivism and the Naturalistic Challenge." In *Routledge Handbook on Liberal Naturalism*, edited by M. De Caro and D. Macarthur. Abingdon: Routledge, chapt. 27.

Bird, G. 1995. "'Kant and Naturalism." *British Journal for the History of Philosophy* 3 (2): 399–408. https://doi.org/10.1080/09608789508570923.

Bird, G. 1996. "McDowell's Kant: Mind and World." *Philosophy* 71 (276): 219–243. https://doi.org/10.1017/S0031819100041450.

Bird, G. 1998. "Kantian Themes in Contemporary Philosophy." *Aristotelian Society Supplementary Volume* 72 (1): 111–151.

Boyle, M. 2016. "Additive Theories of Rationality: A Critique." *European Journal of Philosophy* 24 (2): 1–29.

Boyle, M. 2024. "Kant on Consciousness and Self-Consciousness." In *The Oxford Handbook of Kant*, edited by A. Gomes and A. Stephenson, 153–172. Oxford: Oxford University Press.

Conant, J. 2016. "Why Kant Is Not a Kantian." *Philosophical Topics* 44 (1): 75–125. https://doi.org/10.5840/philtopics20164417.

D'Amasio, A. R. 2021. *Feeling & Knowing: Making Minds Conscious*. New York: Pantheon Books.

Dennett, D. 1987. *The Intentional Stance*. Cambridge, MA, and London: MIT Press.
Frierson, P. R. 2014. *Kant's Empirical Psychology*. Cambridge: Cambridge University Press.
Korsgaard, C. M. 1996. *The Sources of Normativity*. Cambridge: Cambridge University Press.
Louden, R. B. 2020. "Kant the Naturalist." *Journal of Transcendental Philosophy* 1:3–17. https://doi.org/10.1515/jtph-2019-0001.
McCarty, R. 2008. "Kant's Incorporation Requirement: Freedom and Character in the Empirical World." *Canadian Journal of Philosophy* 38 (3): 425–451. https://doi.org/10.1353/cjp.0.0025.
McDowell, J. 1979. "Virtue and Reason." *The Monist* 62:331–350. https://doi.org/10.5840/monist197962319.
McDowell, J. 1996. *Mind and World*. Cambridge, MA: Harvard University Press.
McDowell, J. 1998. *"Two Sorts of Naturalism," In* Mind, Value, and Reality, 167–197. Cambridge, MA: Harvard University Press.
McDowell, J. 1999. *Reason and Nature: Lecture and Colloquium in Münster*. Münster: LIT Verlag.
McDowell, J. 2009. *Having the World in View: Essays on Kant, Hegel, and Sellars*. Cambridge, MA: Harvard University Press.
McLear, C. 2022. "Rationality: What Difference Does It Make?" *Philosophy and Phenomenological Research* 107 (1): 99–112. https://doi.org/10.1111/phpr.12905.
O'Neill, O. 1989. *Constructions of Reason: Explorations of Kant's Practical Philosophy*. Cambridge: Cambridge University Press.
O'Neill, O. 1996. *Towards Justice and Virtue: A Constructive Account of Practical Reasoning*. Cambridge: Cambridge University Press.
O'Neill, O. 2004. "Kant: Rationality as Practical Reason." In *The Oxford Handbook of Rationality*, edited by A. R. Mele and P. Rawling, 93–109. Oxford: Oxford University Press.
Pendlebury, T. A. 2021. "The Shape of the Kantian Mind." *Philosophy and Phenomenological Research* 104 (2): 364–387. https://doi.org/10.1111/phpr.12767 .
Skorupski, J. 1993. *English-Language Philosophy 1750–1945*. Oxford and New York: Oxford University Press.
Stroud, B. 1994. "Kantian Argument, Conceptual Capacities and Vulnerability." In *Kant and Contemporary Epistemology*, edited by P. Parrini and Dordrecht, 231–251. Boston, and London: Kluwer Academic Publishers.
Thompson, M. 2004. "Apprehending Human Form." *Royal Institute of Philosophy Supplement* 54:47–74. https://doi.org/10.1017/S1358246100008444.
Timmermann, J. 2022. Kant's Will at the Crossroads: An Essay on the Failings of Practical Rationality. Oxford: Oxford University Press.
Wood, W. A. 1999. *Kant's Ethical Thought*. Cambridge: Cambridge University Press.