Normative Ontology of Freedom and the Justification of Morality

Igor Tantsorov
ORCID iD: 0009-0000-1899-8719
25.11.2025
This is a non-peer-reviewed preprint

Abstract. The paper proposes a radical rethinking of the foundations of morality by developing a conception of freedom as intrinsically obligatory. In contrast to approaches that derive duty from rationality or reduce it to social conventions, it argues that the normative force of moral requirements arises from the distinctive ontological status of freedom. By distinguishing the "natural" freedom to choose means from the "social" freedom to determine one's ends, the paper shows that the pursuit of the social freedom itself gives rise to universal moral principles. Thus, freedom and morality are interdependent: obligation is a constitutive mode of the free agent's being, and morality is a necessary condition for the realization of freedom.

Keywords: normativity, freedom, constitutivism, moral realism, agency, metaphysics of morality.

Introduction

The relation between freedom and morality presents a well-known set of philosophical challenges. A widely accepted view holds that freedom is a necessary condition for moral responsibility [Fischer, Kane, Pereboom, & Vargas, 2024]. Yet it remains unclear why a free agent should comply with moral norms, often understood as constraints on freedom. Kant notes that "the ability to understand how freedom […] grounds the very possibility of ethical agency lies beyond the reach of theoretical reason" [Kant 2002, p. 65; Ak. 5:46]. This captures the core of the justificatory problem in ethics: the question "why be moral?" has been extensively discussed yet still lacks a satisfactory answer [Himmelmann & Louden, 2015]. In many philosophical traditions, a free agent is assumed to be capable of choosing between morally significant alternatives—for example, in existentialist accounts or in the theistic notion of "moral freedom." Conversely, some approaches

invert the relationship and ask whether morality is a necessary condition for free action. In ancient ethics, as well as in later Platonic traditions, moral action was often identified with genuinely free action [Hecht 2014; Adler 1961]. In contemporary debates about free will [Iredale 2012], however, freedom is typically understood as an ontological rather than an ethical notion [Clarke 2003; List 2019].

These diverse approaches leave one fundamental question unresolved: how can freedom and morality be made conceptually intelligible in terms of each other if neither is reducible to the other? Some approaches derive morality from freedom; others derive freedom from morality; still others separate them entirely. Clarifying the structure of this interdependence requires a more articulated model of agency. The present work develops such a model and uses it to illuminate the conceptual interdependence between freedom and morality.

The aim of this work is not to describe actual social processes but to clarify the conceptual and dynamic interrelations between freedom and morality (understood here as a system of norms). The central claim is that interaction among agents of a certain type necessarily gives rise to universal moral norms, and that adherence to such norms creates the conditions for the freedom of all. Morality thus functions to overcome social—and, indirectly, natural—forms of determination: moral norms do not merely restrict the choice of means for pursuing a predetermined natural (metaphysical) end, but instead nullify that end as a source of determination. In doing so, morality opens a space for self-realization and orientates the agent toward contributing to the freedom of all. Freedom and morality therefore emerge as interdependent in a way that justifies the metaphysical reality of freedom and unifies deontological and consequentialist perspectives

The structure of the paper is as follows. Section 1 introduces natural freedom (N-freedom), understood as the freedom to choose means for pursuing a predetermined natural end. I argue that N-freedom is sufficient for moral responsibility but insufficient for universal moral norms, since such norms conflict with the agent's natural telos. Section 2 introduces social freedom (S-freedom), the capacity to determine one's ends. Section 3 argues that, given the structural features of S-freedom, S-agents have reason to adopt universalizable constraints on action, since only such constraints preserve the reflexive possibility of choosing ends. Section 4 analyzes the compatibility of N-freedom and S-freedom and contends that a society of agents possessing both forms of freedom can function stably only if

additional moral norms are introduced. The final section shows how the introduction of moral norms into a society of N-agents can lead to the emergence of S-freedom.

1 "Natural" Freedom

The prevailing account in contemporary analytic philosophy defines free will as "an agent's ability to exercise the kind of control in action required for moral responsibility" [Caruso & Pereboom 2022]. This approach allows one to analyze freedom in terms of moral responsibility. As Fischer notes, "Some philosophers tend to begin with the notion of moral responsibility and 'work back' to a notion of freedom. For such philosophers, 'freedom' refers to whatever conditions are involved in choosing or acting in such a way as to be morally responsible" [Fischer 2005, p. xxiii]. If this approach is correct, then the presence of a certain internal capacity for control is a necessary condition for morality—understood as a system of norms tied to responsibility for their violation—to arise at all. This thesis finds support in a number of authors.[1]

However, how exactly does freedom lead to the emergence of universal moral norms? To answer this question, we must first clarify what these abilities amount to. Leaving aside external conditions, consider, then, the *minimal* self-control capacities required for an agent to be reasonably regarded as morally responsible (schematically: (i) a cause of (ii) the violating of (iii) a norm).

(i) **Attributability**. An agent must be capable of initiating its behavior in accordance with its own intentions, desires, and motives that arise from its *nature* rather than being imposed externally. In the terms of Chisholm's agent-causal theory, this means that the subject serves as the primary cause of the action rather than as merely a link in an event-causal chain [Chisholm 1964]. Such self-determination is sufficient for attributing an action to the agent, but it is practically impossible without purposiveness. Here "nature" denotes the agent's inner essence that grounds its internal motives; the origin of this nature—genes, upbringing, environment—is irrelevant in the present context. The historical dimension of agency is omitted for two reasons: (1) it is generally intractable, since an agent's nature lies outside its full control and thus "complete" self-determination cannot be required; and (2) it is unnecessary for the minimally sufficient conditions. For a recent discussion, see Ke Zhang [forthcoming].

(ii) **Accountability**. An agent must be able to control its intentions in accordance with external requirements—that is, possess the capacity to learn and adapt. Such control presupposes an ability to modify intentions. These capacities bring the description close to the deep-self / real-self family of views [Frankfurt, 1971; Watson, 1975; Wolf, 1990], but they are not identical to them: modifying intentions (first-order volitions) in response to external demands is not equivalent to controlling one's motives (second-order volitions), let alone doing so in a way that aligns with the agent's moral motivations. Rather, external demands function as a situational, *artificial substitute* for natural second-order volitions: the agent complies without altering its nature. Likewise, capacity (ii) should not be identified with reason-responsiveness, since the agent's nature need not be rational. Because such modification does not affect the agent's nature, this form of control does not imply guaranteed compliance.[2]

(iii) **Moral accountability**. An agent must be capable of understanding a requirement expressed as a formal moral norm; that is, it must be rational enough to learn the content of the norm and to relate it to its actions. Such minimal intelligence ("rational self-control") functionally corresponds to "responsiveness to reasons" [Fischer & Ravizza, 1998] or to "normative competence" [Wallace, 1994]. Rationality, however, does not imply altering the agent's nature in the direction of moral motivation (for example, psychopaths may cognitively understand norms, and this may suffice for holding them responsible).[3]

The type of freedom provided by these capacities may be called "*natural*" freedom[4] (N-freedom), and the corresponding model the N-agent: under conditions of normal development, the agent possesses it *by default* and requires no effort to exercise it. Likewise, under normal conditions (i.e., in the absence of unsolicited interference), N-freedom requires no care—neither expansion nor preservation—it is simply always present. Since physical determinism, whether true or not, does not constrain N-freedom, the latter is fully compatible with it.

A fundamental feature of N-freedom is its limitedness. The N-agent is determined by its nature and by the external conditions to which it must adapt. (Accordingly, the difference between good and evil N-agents is a matter of moral luck.) The unalterable nature of the agent is its essential characteristic, and among the constraints it imposes, the *constitutive* one is the predetermined final end (*telos*), which we may call "well-being." The agent also has a predetermined set of means for

achieving this end. Thus N-freedom can be understood as a relative freedom to choose among these means.

N-freedom aligns naturally with rationality. Because the final end is necessary, transparent, and requires no further justification, an N-agent can in principle possess full practical rationality—both structural and substantive: it can always provide a convincing explanation of its actions, including its compliance with moral norms.[5] Despite this, combining N-freedom with moral norms is problematic. Such norms (i) do not arise from the agent's nature and are therefore "unnatural," and (ii) presuppose voluntariness. Because moral norms introduce an additional externally imposed limitation on the probability of achieving the final end, there is no rational way to justify an unconditional obligation to follow them [see Fehige & Wessels, 2021, for a critique of deriving "ought" from rationality]. Whatever rational considerations may be offered in their favor—social benefit or long-term self-interest—an N-agent retains not only the freedom but also the practical reasons to refuse compliance [see Rational Choice Theory in Lovett & Frank, 2006]. The most reliable mechanism for securing compliance by an N-agent is coercion—that is, turning obligation into necessity. Thus neither the purpose nor the moral meaning of norms is required for responsibility or for normative behavior; understanding its personal consequences suffices.

How could moral norms arise in such an environment? Imagine a society consisting exclusively of N-agents. Let it be sufficiently large that personal relationships can be ignored and only anonymous interactions need to be considered.[6] In such a setting, each agent, in pursuing its predetermined end, inevitably comes into conflict with others: the N-freedom of one agent is incompatible with the N-freedom of any other. Recognizing that all agents are alike, each views the others as competitors, which precludes trust. Under these circumstances, the most advantageous strategy is to exploit any available advantage to advance one's interests ["defection" in Axelrod, 1984; see also Aumann, 2006]. This highlights the natural inequality among agents who possess different qualities relevant to achieving their ends. As a result, even the slightest possibility for the idea of moral equality disappears. Moreover, moral equality would require fair consideration of others' prudential interests, which conflicts with moral intuition [Rawls, 1971, pp. 31, 564]. Accordingly, neither social-contract theory nor game-theoretic models of cooperation apply to such a society [Franklin, 2023; Verbeek &

Morris, 2010].[7] Instead of agreement and cooperation, N-agents resort to forceful domination. Universal and just moral norms cannot arise in such a society: equality as a moral principle cannot be derived from individual rationality or from the structure of interactions under such conditions [Gaus, 2011]. In more realistic terms, such a society inevitably forms hierarchical structures—together with correspondingly unjust norms—based on strength, domination, access to resources, or specialized skills. This hierarchical character is a direct consequence of the natural inequality among N-agents.

Thus "natural" freedom is insufficient for the emergence of just and universal moral norms. Intuitively, such norms presuppose not only the moral equality of agents but also at least a prohibition on causing unjustified harm. Recognizing the necessity of such norms—and being able to follow them without external coercion —requires additional capacities for self-governance that N-agents lack. Let us call the freedom provided by these capacities "*social*" freedom (S-freedom), and the corresponding model the S-agent. (It may not be the most suitable term, since "social freedom" already has established uses in social and political philosophy. Alternatives might include "real," "ultimate," or "authentic.")

2 "Social" Freedom

Let us clarify what S-freedom consists in. Its most evident difference from N-freedom lies in the structure of an agent's ends: only this difference can explain the development of norms that restrict the pursuit of the predetermined natural end and the voluntary compliance with such norms. Because the natural end cannot simply disappear, the social agent must have other, more highly prioritized ends. Several conclusions follow.

First: we are dealing with a phenomenon at least as fundamental in its ontological level as biological determination.

Second. Even when freed from the predetermined end, the S-agent retains its nature. (From now on, we shall understand an S-agent as an agent that has succeeded in liberating itself fully at least from the predetermined end.) Because both its nature and the surrounding environment are structured for that end [Garson, 2017] and not for any other, the S-agent finds itself in an asymmetric situation: it possesses S-freedom yet lacks the possibility of fully realizing it. Conceptually, the agent is continuously positioned to exercise its agency and

therefore—unlike the N-agent—encounters external constraints regardless of the presence of other agents. Accordingly, the S-agent experiences this asymmetry as a kind of pressure urging it to divest itself of S-freedom. Since possessing capacities is not equivalent to exercising them, being an S-agent is not ontologically inescapable [Ferrero, 2009]. It follows that, in order to remain an S-agent under such conditions, it must possess a motivating "force" that *drives* it to employ these capacities. The most plausible source of this force is S-freedom itself.

Third. S-freedom, unlike N-freedom, includes the freedom to choose not only means but also ends. Yet from what, and how, is an S-agent to choose an end? The enigma of ends generates an enigma of means: from where would an S-agent obtain the means needed to pursue its chosen ends? One might attempt to clarify the situation by appealing to human practice. However, once we set aside ends tied to natural determination, the only positive content of personal freedom (Berlin's "freedom to") is self-realization, which ultimately reduces again to choosing one's own ends and realizing one's own capacities. Alternatively, one might try to identify particular obstacles confronting a human being ("freedom from"): an obstacle suggests a direction of movement and thus an end. Yet *anything* may become an obstacle to someone at some time—from foreign ideas to one's own life. Therefore, obstacles cannot provide a reliable explanation for end-selection.[8]

It is therefore no exaggeration to conclude that the aim of S-freedom is S-freedom itself: the possibilities it generates through the overcoming of obstacles are merely possibilities for further freedom. In other words, S-freedom possesses a reflexive teleology: it aims to maintain and reproduce the very conditions of its own realization.[9] The only satisfactory explanation of these facts is that, unlike instrumental N-freedom, S-freedom is a fully self-sufficient phenomenon. We are thus compelled to attribute to S-freedom a set of distinctive and rather unusual properties:

(i) In its *negative* interpretation: since S-freedom is not embedded in the natural order and is not determined by natural teleology, it is not constrained by the forms of determination characteristic of the natural world. It is freedom from any restriction and any necessity. For an S-agent, this manifests as an "insufficiency" of freedom: it always demands further expansion and is therefore potentially unbounded.[10]

(ii) Because S-freedom is not governed by natural laws and possesses a self-

grounding structure—including both the source of its own activity and its own aim—it is not amenable to exhaustive theoretical analysis. Its epistemic status is therefore fundamentally problematic.[11] Thus any definition of S-freedom would simultaneously constitute a limitation and hence a distortion of it.[12] In attempting to understand S-freedom, one must rely primarily on intuition; rationality and empirical experience play only subordinate roles. Accordingly, it is in principle impossible to provide a substantive (*positive*) account of S-freedom.

(iii) The ontological status of S-freedom is likewise problematic: since everything that exists (and everything thinkable) functions as its limitation, we cannot say that it exists in the usual sense; yet we also cannot say that it does not exist, for only the assumption of S-freedom renders the very fact of its limitation intelligible. Thus S-freedom must be conceived as something standing outside the order of beings and therefore not subject to theoretical confirmation or refutation. (That is, it possesses not only epistemic incompleteness but also ontological indeterminacy.) Its ontological status must therefore be accepted as a matter of faith and, within the theory, treated as a *metaphysical postulate*.

About the social agent:

(i) An agent that theoretically possesses full S-freedom has no fixed characteristics other than freedom itself; in this sense, it becomes a pure abstraction. But this does not mean that such an agent disappears. On the contrary: possessing unlimited possibilities, it is free from any stable identity—it can, structurally speaking, become anyone or anything at any time. Hence such an agent is ultimately unknowable (cognitively opaque). Only its constitutive principles are fully knowable.

(ii) Because complete S-freedom is unattainable, striving for it is a *constitutive* principle of the S-agent's behavior—in other words, striving for S-freedom is inseparable from S-freedom itself. Let us call these mutually determining conditions the "*principle of freedom*": an agent strives for freedom insofar as it is free, and is free insofar as it strives for freedom. One may draw an analogy with living matter: if "to be alive" is equivalent to "to strive to be alive," then "to be free" is equivalent to "to strive to be free."

(iii) Although an S-agent can overcome any obstacle, it cannot "free itself" from freedom itself (without thereby ceasing to be an S-agent). Accordingly, since striving for freedom is not a necessity, it becomes a kind of obligation or duty for the S-

agent. That is, unlike the necessity characteristic of determinism, S-freedom possesses a modality similar in appearance yet different in essence—*obligatoriness*: a motivation not implanted in the agent from outside (by nature) but initiated by the agent itself. Continuing the analogy with life: whereas "striving to be alive" is necessary but not obligatory, "striving to be free" is not necessary but obligatory.

(iv) Overcoming obstacles requires effort, and since S-freedom confronts everything that exists, it is impossible without such effort. *Will* is the capacity for this effort. The strength of will is, of course, irregular—an agent can hardly exercise it constantly or uniformly—and therefore cannot be rule-bound. Will is motivated by S-freedom, directed toward it, and—like S-freedom—is obligatory for the S-agent (motivated by duty, in Kant's sense). Thus, unlike N-freedom, which is "given" to the human being, S-freedom is not only the source but also the *product* of volitional effort.

(v) Endowing these volitional efforts with meaning and thereby giving positive content to the S-freedom attained is a creative task of the S-agent. This is the idea of self-realization: the S-agent creates its own being ("itself"), a task undetermined by nature or by the laws of reason. [13] In this way S-freedom manifests as the capacity to generate the new. Thus, in practice, S-freedom is freedom *for*—and *through*—creativity, whose products constitute further possibilities for S-freedom.

(vi) The unknowability of S-freedom limits the agent's ability to understand its own agency, and S-freedom is "felt" only through the presence or absence of obstacles.[14] Therefore, S-freedom does not generate *rational* reasons to pursue it, and the creative ends and actions of the S-agent are not subject to rational evaluation. Its unknowability prevents S-freedom from functioning as a value usable in rational assessment (as required, for example, in Value-Based Theory). Thus, lacking a substantive dimension, the practical rationality of an S-agent is not identical to that of an N-agent: actions the S-agent cannot fully justify are explained by "higher considerations," "rightness," or their self-evident necessity. Nevertheless, while remaining *reasonable* [see Rawls's distinction between the reasonable and the rational, 1993], the S-agent retains the capacity for instrumental rationality; but since its application is now a free choice, refusing rational action in favor of what is obligatory is not necessarily irrational (unlike for the N-agent; cf. n5). As will be shown later, in such cases the agent relies on *moral* reasons.

Overall,, although S-freedom is not exhaustively definable, for the purposes of

further analysis we adopt the following postulate as its minimal structural characterization: S-freedom is the constitutive capacity of the S-agent to govern ends and the means for achieving them—that is, to invent, revise, and revoke them.

3 The Emergence of Moral Norms

How does S-freedom give rise to moral norms? Let us conduct a thought experiment. Imagine a society consisting of S-agents in which, just as in the natural N-society, no moral norms or personal relationships exist. (The term "society" emphasizes that S-agents interact and therefore possess the requisite cognitive and communicative capacities: they can think and communicate while recognizing others as S-agents.) The mere fact of coexistence does not cancel the principle of S-freedom; thus, the S-freedom of each S-agent will sooner or later come into conflict with the S-freedom of every other S-agent. Given that an S-agent is impossible outside society, we must conclude that restrictions on its S-freedom are inevitable—this is a structural necessity. However, the nature of S-agents precludes the use of conflict-resolution methods characteristic of an N-society.

Equality. The unknowability of S-freedom (its cognitive opacity), combined with the creative methods by which an S-agent attains its ends, rules out any advance knowledge of the potential advantages of each agent. This generates the recognition of their equality as potential parties to conflict.[15] At the same time, it is publicly known that all members of the society are constitutionally alike and all strive for maximal S-freedom. Under these conditions, the presuppositions for unequal restrictions of freedom lose all justification. It follows that S-agents inevitably treat one another in accordance with the *principle of equality*; and since this equality has no natural foundation, it consists solely in equality in freedom, with differences in everything else—or, equivalently, in equality of restrictions on freedom.[16]

Voluntariness. Unlike adaptive N-agents, an S-agent finds violence toward other S-agents unacceptable. The reason is that, being an obstacle to S-freedom, violence provokes an endless volitional counteraction, the result of which is the destruction—as an S-agent—of either one party or both once their strength is exhausted. Three outcomes therefore follow:

(i) the S-society becomes empty;

(ii) only one S-agent with the greatest strength of will remains;

(iii) violence, as a form of determination, is rejected by all S-agents.

Since the first two outcomes annihilate society as such, and since both S-agency and S-freedom are possible only within society, outcome (iii) is the only viable possibility. The fact that violence is unacceptable is expressed as the *principle of voluntariness*: an S-agent—and only an S-agent—is free to restrict its own freedom. Self-restriction is possible because an S-agent is capable of adopting any ends, including those that do not conflict with the ends of others.

These two principles follow necessarily from the principle of freedom and mutually sustain one another. Equality is accepted because agents are capable of imposing restrictions on themselves (thereby *equating* themselves with others), and voluntariness is possible because equality makes coercion *universally* rejected. Because its own agency remains opaque to itself, an S-agent experiences these principles—and their violations—primarily in phenomenological form: violence is felt as humiliation, and inequality as injustice. These intuitive reactions are not psychological facts but reflections of the agent's constitutive structure.

Voluntary and equal self-restriction cannot succeed if carried out autonomously, relying solely on one's own understanding of which restrictions on S-freedom are correct, i.e., compatible with the principles of equality and voluntariness. The reason is that such correct restrictions are inaccessible to an autonomous agent: S-freedom allows each agent to understand it in its own way. In other words, what counts as a restriction for one agent need not be perceived as such by another, whereas equality presupposes a criterion independent of the subject. Thus, S-freedom can be restricted only on the basis of explicitly agreed-upon rules that embody such a criterion, which means that S-freedom cannot be realized individually but only *collectively*.[17] From the principle of equality it also follows that common rules must:

(i) apply to all S-agents, i.e. be *universal*;

(ii) apply equally, i.e. be formulated *impartially*.[18]

Where do the rules come from? The function of rules is to eliminate obstacles that arise as consequences of agents' actions; thus the required restrictions must take the form of prohibitions of such actions. But these "social" obstacles—unlike natural forces, which act on the senses and display causal relations—determine agents structurally and manifest themselves only indirectly, and therefore are not always empirically observable or theoretically analyzable (for example, when they restrict future possibilities). Hence, in general, such prohibiting rules cannot be derived by strict scientific methods. This necessitates the use of heuristics and, in

turn, gives rise to a *collective practice*: jointly analyzing real actions that create obstacles and constructing—developing and testing—rules that prohibit them.

Accordingly, the emergence of moral norms takes the form of an *evolutionary* process (in the functional sense). The need for rules—and even for the process itself—is initially unrecognized. Its origin is indistinguishable from ordinary interactions between agents, and the earliest norms appear as customs or traditions. Over time, the process becomes more explicit, and prohibitions come to be understood as rules. As the situation becomes clearer, the process can be formalized and take the form of social dialogue or agreement, and the rules can acquire the status of laws. Since bad-faith participation in the process is indistinguishable from other potential conflicts, norms extend to the process itself, making it recursive. Likewise, non-participation (and the resulting ignorance of the norms) generates future conflicts, leading to requirements of generality (normative consensus).

Because S-freedom is boundless, the process has no natural endpoint. Since agents prefer minimal self-restrictions, the process begins with a single norm but leads to a continual expansion of norms, because each norm further extends the domain of shared S-freedom. At the same time, because new norms apply to the process itself, the process improves as it develops, and the norms it produces improve accordingly. This makes it reasonable to expect the existence of ideal (universal and impartial) rules toward which practical norms recursively converge. By eliminating the social obstacles at which the process was initially aimed, practical norms free agents from any restrictions that would make the success of the process impossible.

The core principle underlying moral norms consists in prohibiting any forms of non-voluntary restriction of S-freedom caused by the actions of S-agents [Svobodin, 2014]. These are precisely the actions that lead to conflict. In general terms, they can be called "violence," and their result—the non-voluntary restriction of freedom—"harm." Clearly, violence and harm may take any conceivable form; precise definitions are impossible because S-freedom is itself undefinable. Every new basic prohibition is, in essence, a further specification of these concepts—meaning that the process itself is a procedure of definition: the content of the terms is interactively and recursively refined in the course of the process. The openness of the terms reflects the dynamic character of morality. Both the principle of "*prohibiting violence*" and the principle of "*prohibiting harm*" are merely alternative

formulations of the principle of voluntariness, expressible also in the form of a maxim: "*Do not restrict the freedom of others*." In this formulation, it functions both as a universal principle for constructing moral norms and as a principle of conduct that helps to avoid conflict. Practical norms based on this principle may be called *just* norms, in contrast to the hierarchical norms characteristic of an N-society.

The fact that norms are followed voluntarily means that intentional violations are impossible; therefore moral responsibility in an S-society does not include punishment. To preserve the effectiveness of norms in case of agents' mistakes, reproach is sufficient.

Thus, the thought experiment shows that S-freedom, as a feature of interacting S-agents:

(i) Is compatible only with equally acceptable moral norms voluntarily imposed by all S-agents on themselves. These norms are just and function as a practical approximation of ideal impartial and universal norms.

(ii) Is an obligatory condition—analogous to a "necessary and sufficient" condition—for the emergence of just and universal moral norms. Because S-freedom, by its nature, cannot determine anything, the appearance of norms at a particular moment is not guaranteed: the construction of norms remains an unpredictable process. (One may say that moral norms are weakly emergent properties of the system: they arise from—but cannot be reduced to—S-freedom.)

(iii) Becomes the source of a specific type of *non-rational* practical reasons—namely, *moral* reasons. These reasons demand unconditional adherence to moral norms. In contrast to the rational reasons of an N-agent responding to external demands out of necessity, the S-agent's response to moral reasons is not necessary but obligatory.[19]

4 The Stability of Moral Norms

Let us now imagine an analogous society, but one that already possesses a set of correct moral norms and consists not of ideal S-agents but of more realistic, finite N+S-agents who possess both N-freedom and S-freedom. This means that the motives of an N+S-agent are simultaneously directed toward S-freedom and toward its own prudential well-being; that is, the N+S-agent is partly free and partly determined by the pre-given end, and both aspects constitute independent sources of motivation.

It is evident that from the standpoint of S-freedom, determination is a limitation, and therefore part of the agent's volitional effort must be devoted to overcoming it. This may lead to a weakening of the striving for S-freedom or even to its temporary disappearance due to fatigue (weakness of will). In such a case, the agent functionally loses S-freedom and reverts to N-freedom.

The existence of these two alternatives may create the *illusion* of the possibility of a "free" choice between them, but such a choice is impossible: neither N-agency nor S-agency is optional for an N+S-agent, in the sense that the agent cannot divest itself of either. The former is necessary, and the latter is obligatory.[20] The apparent ability to choose arbitrarily between what one "wants" (N-motives) and what is "right" (S-motives) is merely a temptation to passive submission to a determining motive, not a free, conscious, and deliberative act. An N+S-agent either manifests sufficient strength of will or succumbs to natural determination.[21]

The determination of an N+S-agent by the pre-given end is not reducible to simple indulgence in desires or passions. A prudential motive, as in the case of an N-agent, may be fully rationally justified. Moreover, rejecting such a motive in favor of a moral obligation renders the N+S-agent even less rational (constitutionally irrational) than an S-agent, for the N+S-agent now rejects fully rational reasons grounded in necessity in favor of what cannot be rationally explained. On the other hand, the N+S-agent is capable of evaluating prudential motives from the standpoint of morality. Yet such reflection does not create a choice, since it does not alter the agent's constitutive motives, but only shapes subsequent self-evaluation—for example, producing shame in cases of weakness of will or pride in the opposite case (or vice versa).

The benefit of such reflection is different. S-freedom requires continuous volitional effort, whereas the N+S-agent's strength is, naturally, limited. Accordingly, periodic reversion to N-freedom—i.e., to internal determination—is practically inevitable. Thus, the N+S-agent's reflective evaluation of its motives is not the justification of *choosing* natural desires, as it may appear, but an assessment of the degree of necessity or desirability of *yielding* to them. Ultimately, deliberation leads to the formation of a pattern of conduct governing the allocation of a finite supply of volitional resources; this allocation is itself a mode of realizing the striving for S-freedom.[22] In the resulting distribution of volitional effort—reflecting both moral and prudential assessments of motives—there appears what may be called the N+S-

agent's "moral character." This shows that such "planning" is not purely rational: the manifestation of character is *individual* and reflects personal strength of will, not norms of rationality as such.

Right action is constrained by moral norms: periodic returns to N-freedom must not extend beyond their limits. Norms may therefore be viewed as the formal boundary between the portion of personal N-freedom that remains to the agent and the already established portion of shared S-freedom—which is, at minimum, limited in those areas where the actions of S-agents and N-agents would otherwise coincide. From this it follows that among freely chosen final ends there is *no place* for the pre-given end. Since the latter does not disappear, it becomes instrumental—it turns into a means. Following this logic, one may expect that as moral evolution proceeds, fewer freely chosen ends will remain that can cause harm to others.

Weakness of will has consequences for morality. First, some agents will deliberately violate norms. Although such behavior is rationally justifiable, its true cause is simply insufficient willpower. The more such weak agents there are, the less S-freedom remains within the society (the more obstacles arise), and therefore the greater the willpower required of each member. Thus, the existence of effective moral norms in an N+S-society depends both on a sufficiently high degree of willpower among N+S-agents and on a sufficient number of such "law-abiding" agents. Moreover, once their proportion falls below a certain critical threshold, the efficacy of the norms collapses entirely, for just moral norms restrict N-freedom and thereby place those who comply with them at a disadvantage, making them victims of the others.

This deterministic process of societal breakdown is not an ordinary social obstacle to S-freedom, because violations of prohibitions cannot be corrected by introducing further prohibitions. Counteracting it requires measures aimed specifically at maintaining the stability of moral norms—namely, the imposition of additional responsibility for their violation. Since violations may also result from error, the degree of responsibility varies (although, given the character of S-freedom, no reliable method exists for determining the underlying cause): if the violation is accidental, reproach suffices; otherwise, punishment must be applied. In the case of punishment, the significance of responsibility for an N+S-agent is twofold. Because the agent possesses N-freedom, punishment compels compliance with norms; and because the agent also possesses the beginnings of S-freedom,

punishment stimulates volitional effort, helping transform the agent into a law-abiding N+S-agent.[23] For both these reasons, punishment is appropriate and justified in a *forward-looking* sense [Caruso & Pereboom, 2022]. (A possible reason for the absence of backward-looking responsibility is the absence of personal relationships.)

Unlike reproach, punishment requires concrete action, and therefore its necessity gives rise to *positive* moral requirements—injunctions—corresponding to the previously introduced *negative* requirements, i.e., prohibitions. Since any violation affects shared S-freedom, it concerns everyone; thus injunctions express the obligation of all participants to sustain the stability of norms.[24] Accordingly, the injunction, like the principle of voluntariness, may be formulated as a maxim: "*Do not tolerate violations of moral norms*."

A second consequence of weakness of will concerns the process of moral evolution. First, punishment requires the formalization of norms and their transformation into public rules or laws; thus the process acquires an explicit and even institutional character earlier than it otherwise would. Second, not all norms easily obeyed by S-agents are equally easy for N+S-agents, who must struggle against natural determination. Third, additional challenges of mutual understanding and agreement arise due to the presence of prudential motives. All of this complicates the process. However, if violators are excluded from participation as part of their punishment, these difficulties are not fundamental and merely slow the process down; if violators do participate, the norms governing the process itself will be violated, thereby threatening the "convergence" of moral evolution. The question of the origin of the process will be addressed later.

Since without appropriate punishment norms could scarcely become entrenched in an N+S-society, it is clear that the requirement to follow injunctions is as much a result of moral evolution as the requirement not to violate prohibitions. This means that despite their functional differences, their moral status is *identical*. Even a stable N+S-society is not "well-ordered" in the Hobbesian sense: punishment is never guaranteed by an external mechanism and always depends on the volitional efforts of the agents themselves. Thus, an N+S-agent always has a rational incentive to violate norms, and therefore both types of moral requirements are structurally inseparable. It follows that an N+S-agent is morally responsible not only for actions but also for its failures to act.

The vice of neglecting injunctions (passivity, conformity) is a sign of weakness of will; however, there are technical difficulties in correcting this vice through punishment. First, fulfilling a positive requirement is relatively more difficult than fulfilling a negative one (it is easier to refrain than to act), which potentially requires greater punishment in order to stimulate compliance. Second, it is harder to establish the fact and degree of violating a positive requirement: the range of possible ways to fulfill an injunction is wide, reducing the feasibility of punishment. Therefore, reproach is typically a more practical means of motivating agents to fulfill positive requirements, and only occasionally should neglect of injunctions be punished (e.g., in cases of complicity). In addition, intolerance toward a violation often implies a conflict with the violator, which—given a general aversion to violence—makes fulfilling this positive requirement more psychologically and socially difficult than obeying a prohibition. This practical asymmetry can produce the illusion that injunctions are secondary to prohibitions: as is sometimes argued, "it is better to allow harm than to cause it oneself" [Foot, 1967]. However, "easier" does not mean "better": the practical difficulty of fulfilling an obligation does not determine its moral status.[25]

Thus, the more realistic model of the agent reveals no fundamental obstacles to the stability of moral norms under two conditions:

(i) a sufficiently small number of weak-willed N+S-agents;

(ii) the presence of appropriate positive moral requirements.

5 The Emergence of Freedom

Let us return to moral evolution. An N-society is incapable of generating either just norms or S-freedom. Where, then, can they come from? As for norms, they could in principle arise *ex nihilo*. Suppose they were proposed by "philosophers" on the basis of thought experiments. (This does not contradict the absence of S-freedom, which is not required for free thinking.) Of course, without practical testing and collective deliberation, the quality of the proposed norms would be mediocre, but someone must first propose norms for them to be collectively evaluated. Accepting this hypothetical assumption, we may now attempt to imagine the emergence of S-freedom.

Imagine a society composed of N-agents into which just moral norms have been introduced. It is clear that such an experiment is doomed: N-agents, if they obey the

new requirements at all, will do so only under coercion. But let us further suppose that among N-agents there are some who possess a latent capacity—not as a psychological trait but as an ontological possibility of belonging to a distinct metaphysical order —for S-freedom, a capacity left unactualized under N-society conditions. In other words, an N-society contains potential N+S-agents. This assumption, like the first, is quite plausible, since the source of the striving for S-freedom is unknown and can be logically assumed to exist in at least some N-agents.

At this point a certain dynamic appears: moral requirements help a potential N+S-agent become an actual one. All that is needed is the voluntary acceptance of a norm as a basis for action. This occurs because the obligations implied by moral requirements outline the contours of a future and thereby provide a hint of possible ends. By restricting N-freedom, morality guides the agent, helping it select other, correct ends.[26] Of course, the pre-given natural end is not directly prohibited; however, the fact that the agent voluntarily limits itself means that the pre-given end loses its determining force. Voluntariness is the manifestation of S-freedom: it has become the agent's ultimate end, while the pre-given end—perhaps implicitly—has transformed from an end into a means.

But for this transformation to become real—for an N-agent not merely to be swept along by trends or whims, but genuinely to exert will and renounce the old end—it must recognize the moral correctness of the norms.[27] And to do that, it must "feel" S-freedom within itself and recognize itself as an (N+)S-agent—something that is not problematic if the seeds of S-freedom are already present. S-freedom discloses within the agent the capacity for moral evaluation (and for self-evaluation; see n3), and this disclosure is made possible by the objective fairness of the new moral norms, especially by their contrast with humiliating or unjust ones. Self-evaluation reveals to the agent its moral equality with others, and equality enables normative communication (which identifies other such agents). Communication generates mutual trust, moral support, joint acceptance of moral norms, and subsequently moral pressure on hesitant members of the society.[28] Under favorable circumstances, this dynamic is sufficient to initiate the transformation of an N-society into an N+S-society and to sustain its further evolution.

The described transformation of a potential N+S-agent into an actual one inevitably confronts the agent with a question of meaning: *why be moral*? The only possible answer—"to be free"—is not rational in the sense of offering motivating

reasons. There is neither necessity nor possibility for an N-agent to overcome its determinacy and become free. The answer "to be free" is instead a statement of the fact of S-freedom and reflects the essential feature of an S-agent (or an actualized N+S-agent): it cannot fail to be free and thus cannot fail to be moral [cf. Korsgaard 1996].

Yet the search for meaning may further raise the question: *why be free*? Although S-freedom requires neither causes nor reasons, it at least allows an agent to have individually determined ends, and thus a sense of meaning. Yet S-freedom guarantees neither the rightness nor the attainability of these ends; it merely makes them genuinely one's own [cf. Wolf 2010]. In contrast, N-freedom—being determined primarily by its end—guarantees the *absence* of meaning: an N-agent simply obeys necessity.

Nevertheless, the ontological indeterminacy (unknowability) of S-freedom shifts the question of meaning to a different register: *where is meaning to be found?* Unknowability provides the answer: an end is understood in the movement toward it, for to be free is to strive for freedom. In other words, meaning is tied to creating the conditions for becoming free; since nature offers no such paths, the task of liberation is creative—ways must be *invented* and *realized*. Moreover, because S-freedom can only be *shared*, individual creative accomplishment acquires meaning only insofar as it expands possibilities for everyone. Thus meaning becomes linked to usefulness to the common cause, and individual self-realization is achieved through one's contribution to the collective movement toward S-freedom—a movement possible only as a form of joint creative labor. Hence the behavioral maxim expressing the creative aspect of S-freedom: "*Create possibilities for everyone to become freer*." Yet the moral status of this principle differs from the previous ones: since creativity is unpredictable, reproaching an agent for failure is hardly appropriate.

Within the unfolding process of moral evolution, the principles of equality, voluntariness, and creativity (including the invention of norms) operate together. S-freedom emerges at the social level because N+S-agents *mutually and harmoniously* restrict their actions: self-restriction by one agent liberates others, and the self-restriction of others liberates one. Thus arises a normatively structured, conflict-free space in which each agent gains the ability to act according to its creative impulses. In these impulses genuine free choice manifests: the agent now invents its

own ends and the means to achieve them. These ends and means realize S-freedom and, unlike those characteristic of N-freedom, are morally rather than externally constrained—they cannot be directed toward goods that conflict with the common good (that is, they cannot cause harm to others). The resulting social space of shared S-freedom, in turn, supports and strengthens individual S-freedom. Thus, by using the agent's striving for freedom, morality transforms private interest into public interest (see n17), allowing one to resolve the tension between the meaninglessness of finite individual existence and the meaningfulness of the infinite collective movement toward the goal. In this way morality helps not only to establish S-freedom in society but also to fill it with positive content.

Thus the final thought experiment shows that the emergence of S-freedom in an N-society requires two conditions:

(i) the presence of just moral norms;

(ii) the presence of a sufficient number of potential N+S-agents capable of responding to these norms.

Concluding Remarks

Universality. The proposed model of agency is limited to a single society and does not include an analysis of personal or group relations. This idealization is justified, since the aim of the work is not to examine the full diversity of social bonds. Interpersonal and group relations are largely contextual and generate primarily positive norms—care, charity, assistance, and so forth. These positive norms impose additional constraints on S-freedom but do not affect either the principle of freedom or the tension between free and determined ends: they can be understood as an expansion of the notion of "one's own good" to include a wider circle of persons (from personal well-being to the well-being of loved ones or the group). Because these positive norms arise from various natural and social needs, they are, on the one hand, less universal than negative norms (and typically carry lower moral priority [Foot, 1967]), and, on the other hand, they conflict both with those norms and with one another. The idealized model, by contrast, allows one to abstract from context and thereby justify universal moral norms, thus contributing to the resolution of moral conflicts.

Cognitivism. Relying on the admittedly limited possibilities for the theoretical analysis of S-freedom, the theory nevertheless allows for the justification of

universal moral principles. Although moral norms are not deduced from these principles, they can and should be regarded as the outcome of collective social construction on their basis, taking practical constraints and conditions into account. The process of searching for and improving norms, in an ideal perspective, aims at what may be considered moral truth. The norms constructed at each stage of the evolutionary process, though not true, can be regarded as correct. Their correctness is affirmed in two ways. If the process is viewed as a formal moral discourse, a norm receives its legitimate grounding in the spirit of Apel [Apel, 1980] and Habermas [Habermas, 1990]. In this case, the constructed norm possesses the status of normative rightness—a kind of validity of prescriptive statements that rests on discursive norms established *before* the given stage (and not fixed from the outset). If, however, the process is considered in its essence, the initial correctness built into its foundation is justified at each stage through consensus and through Peirce's pragmatic criterion of efficiency [Peirce, 1878].

Constitutivism. The proposed approach is a version of constitutivism, since it derives universal moral principles from the constitutive feature of agency. Its distinguishing feature is that these principles are derived for *interacting* agents (see objections to such an approach in [Sem de Maagt, 2019]). Accordingly, insofar as the principle of freedom is constitutive of the (free) agent, the resulting principles are constitutive of a (free) society.[29] The constitutive feature also explains how there could be such a thing as normativity.[30] The feature that transforms a description ("what is constitutive") into a prescription ("what is normative") is motivation by S-freedom: one cannot in the normative sense be a free agent without striving to be one. Thus, the constitutive feature of S-agency serves less to bind by itself than to transmit the normative force of S-freedom. Accordingly, this version of constitutivism, while providing moral norms with metaphysical grounding, is not itself a version of constructivism.

Realism. The self-evidence and universal obligatoriness of fundamental moral principles demand explanation; in the proposed model, they are treated as consequences of the presence of S-freedom in agents. Their stable role in practical reasoning indicates its *reality*[31] as a condition of possibility for moral reasoning, in a sense akin to a transcendental argument: the ontological status of S-freedom is justified by the fact that without it one cannot explain the self-evidence, universality, or normative force of fundamental moral principles. The unusual ontological status

of freedom—whose essentially teleological mode of "presence" binds 'is' and 'ought'—explains not only the phenomenon of moral normativity but also its analogous strangeness ("queerness," according to Mackie [Mackie, 1977]): being irreducible to facts of the natural world, it requires an equally unique grounding. The presented theory can thus be viewed as a variety of *strong non-naturalist realism*.

Objectivity. At first glance, the assumption of the reality of freedom seems to contradict the earlier conclusion that its existence is a matter of belief. Yet for the objective correctness of the theory, what matters is not what agents believe but how they act. The practical striving for freedom does not depend on the presence or absence of a theoretical belief in it. As for the objectivity of moral requirements, the theory relies on two levels of objectivity: on the one hand, the metaphysical objectivity characteristic of realism (the status of principles), and on the other, the weaker objectivity characteristic of constructivism (the status of evolutionarily developed norms—the contingency of human natures and circumstances would preclude comparable objectivity for their prescriptions for action). At the same time, the theory provides a general normative foundation that ensures a more robust objectivity than a purely constructivist one.[32]

Substantiveness. The paper has shown that the striving for freedom leads to the formation of correct moral principles, and that adherence to these principles, in turn, contributes to the realization of freedom. In other words, "right" actions are those that lead to freedom, and "freedom" is what is produced by right actions. The resolution of this apparent paradox lies in the fact that universal moral principles simultaneously (i) realize freedom and (ii) serve as the basis for qualifying actions as right. Thus, the question "Which is more fundamental: rightness or goodness?" has no rational answer from the agent's perspective. Since S-freedom is the ultimate and universal end, the theory is simultaneously consequentialist and deontological: it not only justifies formal principles but also gives them substantive content, indicating the directions of right action.[33] This conclusion accords with the intuitive sense that in morality the rightness of actions is as important as the rightness of ends.

Given these characteristics, the proposed theory provides a promising framework for further research.

Notes

1. As Strawson (1962) argues, practices of moral responsibility are embedded in the structure of the reactive attitudes that arise in response to the observance or violation of commonly accepted standards. Scanlon (1998) similarly characterizes moral responsibility as answerability to others with respect to the observance of norms. Hlobil likewise contends that the very concept of rule-following implicitly presupposes the possibility of violating the rule, which makes the ascription of responsibility possible (Hlobil, forthcoming, p. 15).

2. This is because external demands always stand in potential conflict with desires that arise from the agent's nature, and the capacity to change intentions is not equivalent to the *desire* to change them. Although the outcome of the conflict between motives and requirements is determined by whichever is stronger, from the perspective of an external observer an agent will always appear capable of "doing otherwise." However, the necessary condition for responsibility is not the mere presence of such a capacity (as some incompatibilists require), but the agent's awareness of it—or its taking itself to be aware of it. In our case, the basis for responsibility is the fact that the agent knows about a feasible requirement (the epistemic condition) yet fails to display sufficient desire to comply.

3. "Rational self-control" should not be confused with what is sometimes called "reflective self-control" [Wallace, 1994]. Reflection involves the critical evaluation and subsequent modification of an agent's nature—that is, the formation of second-order desires. Frankfurt called agents capable of evaluating their motives *persons*: in such cases, the temporary substitution of natural second-order volitions becomes permanent and is fully internalized. Since "rational self-control" does not require this evaluative stance toward one's motives, this type of agent models only one aspect of full personhood. In this connection, it is helpful to recall the difference between acting *in accordance with a rule* and *rule-following* [Hlobil, forthcoming], as well as Kant's distinction between *legality* and *morality*.

4. I use "natural" freedom (and later "social" freedom) stipulatively. See, however, the similar use of "natural freedom" in Adler [1961], *The Idea of Freedom*, vol. I, p. 149. It is also important to distinguish this sense from the political notion of "natural liberty" characteristic of social-contract theory.

5. *Substantive* here means endorsing the final end, which is reasonable in itself because it perfectly correlates with the agent's intrinsic desires. This corresponds to

the prudential conception of rationality found in Hobbes. The agent can also explain why it is instrumentally rational: abandoning the pursuit of its predetermined end would be irrational [cf. Dreier, 2001, on the "why be rational?" problem].

6. Personal relationships complicate the picture because they introduce additional motives—reputational, emotional, or kin-based. These mechanisms can explain the emergence of cooperative behavior in small groups but lose much of their force outside close-knit contexts. Moreover, evolutionary mechanisms account not only for cooperative norms but also for norms of aggression—war, genocide, and similar forms of collective violence.

7. Social-contract theories and standard game-theoretic cooperation models presuppose either repeated interactions, or institutional mechanisms for establishing and enforcing agreements, or the symmetry of agents' status [Gauthier, 1986; Skyrms, 1996]. None of these conditions obtain here. Empirical studies likewise show that stable cooperation requires monitoring and sanctions; without them it quickly collapses even among rational, learning agents [Fehr & Gächter, 2002]. Institutional theories of common-pool resource management further demonstrate that stable norms arise only under special structural rules [Ostrom, 2010].

8. Besides the many forms of determination that constrain human beings—cultural, economic, linguistic, psychological, genetic, etc.—even nomological constraints paradoxically function as obstacles: humans not only struggle with entropy but also seek ways around the limits imposed by physical laws (energy, transportation, communication, etc.). The very logic of scientific and technological progress can be read as a practical expression of the central Enlightenment idea—rational mastery of nature—which Fichte described as a condition for material independence and the realization of a rational agent.

9. The idea that the ultimate end of human beings—or humanity—is freedom, or a closely related concept, appears in a long tradition including Rousseau, Kant, Fichte, Hegel, and Sartre.

10. This is not merely an analogue of Kantian transcendental freedom, which is defined negatively as independence from everything empirical. Unlike it, S-freedom entails independence not only from natural necessity but also from other forms of determining influence, including cultural and logical ones. In this sense, S-freedom can be viewed as a form of expanded negative freedom, irreducible to Hobbes–

Hume–Berlin freedom of action or Adler's circumstantial freedom [Adler, 1961, p. 225], and only partially overlapping with Sartre's radical freedom. If one shifts the emphasis from the agent to the structure of reality—understanding S-freedom not as an individual property but as an ontological condition of the possibility of free action—the concept comes to resemble ideas of German Idealism: in particular, Fichte's idea of freedom as self-positing and Schelling's account of freedom as a fundamental property of the Absolute.

11. Given that freedom is understood as the negation of determinism, the epistemology of freedom may be regarded as a kind of apophatic epistemology of determinism. Yet even the very question of the truth of physical determinism remains unresolved.

12. Fichte likewise argued that freedom cannot have a definition because it stands above all knowledge; in this respect, S-freedom resembles the Platonic and Proclean "One," about which one cannot speak directly.

13. Freedom as a creative principle was emphasized not only by the German Idealists but also by Berdyaev, Sartre, Svobodin, and others.

14. In this case the unknowability of freedom again appears as cognitive opacity, and the question of moral motivation has become traditional [Richardson, 2018]: an S-agent does not fully understand what motivates it—whence comes the desire to be free or to follow morality. Nor is it clear whether its free will is compromised by hidden causes.

15. This is not epistemic information-hiding but constitutive indeterminability—unlike, for example, Rawls's "veil of ignorance." There is simply no way to determine the comparative strength of the parties without direct confrontation.

16. This explains, among other things, why concepts of social or political equality are difficult to formalize essentially, whereas the idea of equal freedom has historically been regarded as fundamental.

17. Thus, unlike the N-agent, the S-agent is intrinsically social, which is close to the conceptions of agency in Fichte and Hegel. Cf. also [Sartre, 1988]. This shows that the question of the subject of S-freedom remains as unclear as S-freedom itself —is it the individual, society, or the individual+society? The latter would explain one of the paradoxes of S-freedom: how can something that opposes *any* limitation limit itself? Continuing the comparison with the N-agent, note that whereas S-freedom is characterized by moral equality, N-freedom highlights inequality: one's

own welfare, like one's own life, is never equal to another's.

18. Cf. Scanlon's approach [Scanlon, 1998], which is inapplicable to N-agents but suitable for S-agents.

19. In other words, moral reasons possess a normative force whose source is S-freedom: norms merely "formalize" the striving for freedom, converting it into specific reasons, but they are not its source. This aligns with reasons internalism [Prinz, 2015]. However, in this case one cannot always posit a *rational* connection between motivation and action, which gives rise to the familiar "puzzle" of moral motivation.

20. If such a "choice" were genuinely possible, it could not be grounded in any reasons and would therefore amount to an act of pure randomness; cf. Frankfurt [1999, pp. 110, 114–115].

21. The core of the illusion lies in the fact that, under the guise of free choice, the N+S-agent *always* "chooses" N-freedom, since the opposite "choice" requires additional volitional effort, and in its absence the agent automatically regresses to N-freedom. However, the obligatoriness intrinsic to freedom conflicts with the idea of freedom as choice: if freedom is obligatory, it is no longer an option; if it remains an option, it cannot be obligatory. This paradox is known as the Reinhold–Sidgwick problem [Bojanowski, 2023]. Its essence is that the refusal of S-freedom is simultaneously determined—because there are reasons for it—and not determined—because it depends entirely on the agent's willpower.

22. This self-reflection may be conceived as the rational management of the distribution of will relative to desires, with reason functioning as an integrating mechanism unifying the two sources of motivation and thereby overcoming the agent's constitutive dualism; see also n3.

23. Since the violator does not yet possess sufficient S-freedom, coercion does not violate the principle of voluntariness. However, the presence of nascent S-freedom requires taking the principle of equality into account, which leads to the requirement of proportionality—as a form of justice—in the degree of punishment.

24. Although both prohibitions and injunctions arise from the relation between S-freedom and the structural conditions of coexistence, the kind of necessity involved differs. Prohibitions follow with constitutive necessity from the principles of voluntariness and equality, whereas injunctions possess only a functional necessity required to uphold them.

25. This illusion dissipates once one considers that injunctions are deeply internalized in human psychology. The need to follow a norm oneself can scarcely be separated from the desire to see the same behavior in others. Emotionally, this manifests in what Strawson called "reactive attitudes" [Strawson, 2008], which, in his view, are *inescapable* and, moreover, tend not to remain internal but become expressed as *demands* [Watson, 2014].

26. The shaping of ends is a classical task of ethics: already in Antiquity, one task of ethics was to indicate to the human being the path toward the right life—that is, toward the right life-ends.

27. For realist interpretations of moral correctness, see [Scanlon 2014; Parfit 2011; Timmermann 2018].

28. While the moral community is thus forming its members, holding N-agents responsible for lacking a striving for S-freedom is unjustified: responsibility presupposes the relevant capacities. However, the cognitive opacity of S-freedom leaves a degree of uncertainty: there always remains the logical possibility that a potential N+S-agent is avoiding the actualization of this capacity.

29. This dialectic can explain the paradox of the subject of freedom, see n17: the individual and society mutually constitute one another, since neither an S-agent nor an S-society can exist independently.

30. The core of normativity lies in *obligation*, not in inescapability. Inescapability —being an agent, having a constitutive end, principle, function, etc.—is insufficient, since an agent can always ask: "Why am I obligated to obey inevitability?" See [Ferrero, 2009; Enoch, 2006; Katsafanas, 2018].

31. The reality of S-freedom here is not understood as belonging to any particular ontological category—material, ideal, or substantial—but solely as *its independence from reason*.

32. Such objectivity often appears to dissolve into intersubjective agreement [Axtell, 2015, p. 190]. But intersubjective agreement requires a common normative foundation; without it, the possibility of constructively grounding norms remains problematic: moral norms may endlessly and circularly fluctuate over time as a result of ongoing deliberation [Bagnoli, 2013, pp. 153–182]. The presence of an objective foundation, combined with the recursive structure of deliberation, allows hope for epistemic convergence toward the very metaphysical objectivity on which the process rests. The essential difference between this robust objectivity and

metaphysical objectivity is that the latter is reachable only asymptotically, at a hypothetical limit of an infinite process.

33. As noted by [Timmermann, 2018] and [Schroeder, forthcoming], attempts to give priority either to principles or to ends distort the understanding of normative foundations, whereas their mutual dependence turns out to be structurally necessary.

References

Adler, M. J. (1961). *The idea of freedom*. Garden City, NY: Doubleday.

Apel, K.-O. (1980). The a priori of the communication community and the foundation of ethics: The problem of a rational foundation of ethics in the scientific age. In *Towards a transformation of philosophy* (pp. 225–300). Routledge.

Aumann, R. J. (2006). War and peace. *PNAS, 103*, 17075–17078.

Axelrod, R. (1984). *The evolution of cooperation*. Basic Books.

Axtell, G. (2015). *Objectivity*. Polity Press.

Bagnoli, C. (2013). Constructivism about practical knowledge. In C. Bagnoli (Ed.), *Constructivism in ethics* (pp. 84–104). Cambridge University Press.

Bojanowski, J. (2023). Reinhold on free will and moral obligation: A Kantian response. In J. Saunders (Ed.), *Freedom after Kant* (pp. 17–32). Routledge.

Caruso, G. D., & Pereboom, D. (2022). *Moral responsibility reconsidered*. Cambridge University Press.

Chisholm, R. M. (1964). Human freedom and the self. *The Lindley Lecture*. University of Kansas.

Clarke, R. (2003). *Libertarian accounts of free will*. Oxford University Press.

Dreier, J. (2001). Humean doubts about categorical imperatives. In E. Millgram (Ed.), *Varieties of practical reasoning* (pp. 27–47). MIT Press.

Enoch, D. (2006). Agency, shmagency: Why normativity won't come from what is constitutive of action. *Philosophical Review, 115*(2), 169–198.

Fehige, C., & Wessels, U. (2021). Rationality and morality. In M. Knauff & W. Spohn (Eds.), *The handbook of rationality* (pp. 682–692). MIT Press.

Fehr, E., & Gächter, S. (2002). Altruistic punishment. *Nature, 415*, 137–140.

Ferrero, L. (2009). Constitutivism and the inescapability of agency. In R. Shafer-Landau (Ed.), *Oxford studies in metaethics* (Vol. 4, pp. 303–333). Oxford University Press.

Fischer, J. M. (Ed.). (2005). *Free will: Critical concepts in philosophy*. Routledge.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge University Press.

Fischer, J. M., Kane, R., Pereboom, D., & Vargas, M. (2024). *Four views on free will* (2nd ed.). Wiley-Blackwell.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review, 5*, 5–15.

Frankfurt, H. G. (1971). Freedom of the will. *Journal of Philosophy, 68*, 5–20.

Frankfurt, H. G. (1999). *Necessity, volition, and love*. Cambridge University Press.

Franklin, J. (2023). Let no-one ignorant of geometry… *Journal of Value Inquiry, 57*(2), 365–384.

Garson, J. (2017). A critical overview. *Biology & Philosophy, 32*, 79–105.

Gaus, G. (2011). *The order of public reason*. Cambridge University Press.

Gauthier, D. (1986). *Morals by agreement*. Clarendon Press.

Habermas, J. (1990). *Moral consciousness and communicative action*. MIT Press.

Hecht, J. (2014). Freedom of the will in Plato and Augustine. *British Journal for the History of Philosophy, 22*(2), 196–216.

Himmelmann, B., & Louden, R. B. (Eds.). (2015). *Why be moral?* De Gruyter.

Hlobil, U. (Forthcoming). Intrinsic responsibility for rule-following. *Topoi*.

Huemer, M. (2008). *Ethical intuitionism*. Palgrave Macmillan.

Iredale, M. (2012). *The problem of free will: A contemporary introduction*. Routledge.

Kant, I. (2002). *Critique of practical reason*. Hackett Publishing.

Katsafanas, P. (2018). Constitutivism about practical reasons. In D. Star (Ed.), *Oxford handbook of reasons and normativity* (pp. 367–393). Oxford University Press.

Korsgaard, C. (1996). *The sources of normativity*. Cambridge University Press.

List, C. (2019). *Why free will is real*. Harvard University Press.

Lovett, F. (2006). Rational choice theory and explanation. *Rationality and Society, 18*(2), 237–272.

Mackie, J. L. (1977). *Ethics: Inventing right and wrong*. Penguin Books.

Morriston, W. (2000). What is so good about moral freedom? *Philosophical Quarterly, 50*(200), 344–358.

Ostrom, E. (2010). *Beyond markets and states*. Nobel Prize Lecture.

Parfit, D. (2011). *On what matters*. Oxford University Press.

Peirce, C. S. (1878). How to make our ideas clear. *Popular Science Monthly, 12*, 286–302.

Prinz, J. (2015). An empirical case for motivational internalism. In G. Björnsson et al. (Eds.), *Motivational internalism* (pp. 61–84). Oxford University Press.

Rawls, J. (1971/1999). *A theory of justice*. Harvard University Press.

Rawls, J. (1993). *Political liberalism*. Columbia University Press.

Richardson, H. S. (2018). Moral reasoning. In *Stanford encyclopedia of philosophy*.

Russell, P., & Deery, O. (2013). *The philosophy of free will*. Oxford University Press.

Sartre, J.-P. (1988). *What is literature?* Harvard University Press.

Scanlon, T. M. (1998). *What we owe to each other*. Harvard University Press.

Scanlon, T. M. (2014). *Being realistic about reasons*. Oxford University Press.

Schroeder, S. A. (Forthcoming). Leaving the concept of deontology behind. In *Oxford handbook of normative ethics*. Oxford University Press.

Shafer-Landau, R. (2005). *Moral realism: A defence*. Oxford University Press.

Skyrms, B. (1996). *Evolution of the social contract*. Cambridge University Press.

Strawson, P. F. (1961). Social morality and individual ideal. *Philosophy, 36*, 1–17.

Strawson, P. F. (2008). Freedom and resentment. In *Freedom and resentment and other essays* (pp. 1–28). Routledge.

Svobodin, I. (2014). Objective ethics. In *Authentic Social Contract: Moral Foundations.* Amazon.

Timmermann, J. (2018). The law and the good. In *Natur und Freiheit* (pp. 675–692). De Gruyter.

Verbeek, B., & Morris, C. (2010). Game theory and ethics. *Stanford encyclopedia of philosophy*.

Wallace, R. J. (1994). *Responsibility and the moral sentiments*. Harvard University Press.

Watson, G. (1975). Free agency. *Journal of Philosophy, 72*, 205–220.

Watson, G. (2014). Peter Strawson on responsibility. In *Oxford studies in agency and responsibility* (Vol. 2, pp. 15–32). Oxford University Press.

Wolf, S. (1990). *Freedom within reason*. Oxford University Press.

Wolf, S. (2010). *Meaning in life and why it matters*. Princeton University Press.

Zhang, K. (Forthcoming). Bringing the deep self back. *Analytic Philosophy*.

de Maagt, S. (2019). It only takes two to tango. *Philosophical Studies, 176*(10), 2767–2783.