Abstract

This paper addresses the longstanding Problem of Action by proposing a unified model—the Action Interface—that integrates the causal, rational, and volitional dimensions of agency. Drawing from thinkers like Davidson, Frankfurt, and Hempel, I argue that current theories often speak past one another because they operate at mismatched levels of analysis. The Action Interface offers a layered framework that segments action into three key pipelines: sensory-perception, planning, and volitional framing. This structure allows us to locate and evaluate the specific features that motivate desires, shape intentions, and govern execution. I defend a formal definition of action as the guided execution of movements caused by an actionable desire, contextualized by a scope established at the moment of acting. Finally, I introduce the Counterfactual Substitution Test as a diagnostic tool for assessing intentionality across complex causal chains.

Introduction

To properly frame the Problem of Action, we must first clarify what is meant by *action*, and more crucially, identify the cognitive and linguistic mechanisms through which such a concept arises. This framing is not a mere preliminary—it is foundational. Without it, we risk miscommunication at the very point where precision is most required. It is far too easy, when reasoning about action, to smuggle in intuitive assumptions from isolated instances of behavior—each miscategorized by virtue of an ill-defined schema. To avoid this, I begin with a definitional scaffold: that *action*—whether physical or conceptual—is the output of a co-evolved system, spanning the coordination of sensation-perception mappings with neuro-muscular execution, under the mediation of logical inference and linguistic convention. From this vantage, action is not simply a discrete event, but an evolved strategy—one shaped, in part, by its utility in enhancing survival and reproductive fitness.

Suppose, then, we accept that the concept of action evolved as a pragmatic heuristic. What, precisely, would justify the development of such a computationally expensive faculty? Notably, the burden of this justification does not fall on the agent alone. From a linguistic perspective, there is

no clear reason to internalize a concept of action for purely solipsistic ends. It is far more plausible to treat action as an interpretive instrument—one calibrated to anticipate, evaluate, and negotiate the behavior of others. That is: action as a construct of *social cognition* (Ackerman et al., 2012). On this view, the primary utility of action-language is third-personal. Yet as social ecosystems grow in complexity, this third-person scaffolding begins to turn inward. The same mechanisms that track and categorize others' behaviors are recursively applied to the self—producing what we might call the *reflective layer* of action. This layer, narrative in structure and moral in tone, rationalizes one's own behavior through a taxonomy of prosocial and antisocial framings.

From this perspective, the philosophical impetus to ground action in causal terms becomes more intelligible. Causal models do not emerge from abstract contemplation alone—they are born from interaction with agents operating in causal environments. The need to *explain* one's actions—to oneself and to others—begets the impulse to link them in sequenced, rule-governed ways. Yet herein lies the central difficulty: any attempt to *formalize* action must contend with the continuous and noisy nature of bodily movement. The category of "action" cannot be reduced to discrete kinematic boundaries; it is not defined by movement alone, but by *guidance*, *intentionality*, and *interpretability in context*. If we are to isolate the moment of acting, it seems that we must find some marker—accessible not only post hoc but during the event itself—by which action may be apprehended as such.

This tension—between abstracted cause and embedded guidance—marks a central fault line within contemporary Action Theory. Frankfurt, for instance, asserts that the mere presence of antecedent causes, even rational ones, fails to account for the agent's active relation to the bodily movement as it unfolds (Frankfurt, 1978). Others, like Davidson, seek to repair this by invoking the alignment of beliefs and desires, yet still lean heavily on prior conditions (Davidson, 1963). These competing accounts are not incommensurable, but rather mismatched in scope. What is needed is a structured framework—one that can mediate between their levels of description. To that end, this paper will proceed in two stages. First, I introduce the Action Interface, formalizing its structure and proposing a test for evaluating causal influence within it. Second, I use this framework to

situate and reconcile the arguments of key theorists, illustrating how their views map onto distinct loci within the interface and, once properly contextualized, can be seen as mutually informative rather than mutually exclusive.

The Action Interface

Conceptual Design Structure

The conceptual seeds of the Action Interface model find early expression in Carl Hempel's 1961 essay *Rational Action*, wherein Hempel attempts to formalize a definition of rational action through the framework of empirical hypothesis and critical appraisal. The first component—the hypothesis—posits that certain reasons motivate action, including "ends the agent sought to attain, and his beliefs about available means of attaining them" (Hempel, 1961). At first glance, this formulation appears to have little to do with rationality *per se*; rather, it functions as a scaffolding device—a way of embedding action within a broad enough evaluative context so that it can, subsequently, be assessed on rational grounds. Indeed, Hempel's language is revealing: terms such as "reasons," "motivations," and "ends" evoke Davidson's causal schema, Anscombe's intentional descriptions, and the teleological frameworks of philosophers like George Wilson and Michael Thomson

Though Hempel himself expresses philosophical modesty regarding the project—"Practically, this is no doubt often the best we can do by way of explaining an action . . . If such theoretical developments should show that the explanatory power of the concept of rational action is in fact rather limited, we will have to accept this philosophically"—I would argue that his account gestures further than he admits. When reframed not as a final theory but as a topological bridge, Hempel's model approaches what I take to be the central challenge of Action Theory: the need to reconcile mechanistic explanation with interpretive coherence. For this reason, it is worthwhile to briefly distill the primary features of Hempel's approach before turning to a more systematic interface.

Central to Hempel's framework is a cluster of interlocking terms, chief among them the *information basis*—a set of sentences denoting the totality of relevant information available to the agent at the moment of action. This is paired with what he calls the *total objective*, a construct that can likewise be described by a set of propositions articulating the desired end-state. To this is appended a further set: descriptions of feasible means likely to realize that end, subtracting any which violate constraining norms. Notably, Hempel acknowledges the existence of actions that "do not consist in attaining a specified end state" (Hempel, 1961), and thus brackets them from analysis—effectively conceding their resistance to incorporation within his model.

He further supplements the account by incorporating models from economics and game theory to illustrate the range of strategies an agent might employ to achieve their ends, indicating the absence of a monolithic description under which an action might be considered rational. Lastly, Hempel characterizes *rationality* not as a set of criteria but as a *disposition*—drawing an analogy to the property of magnetism. To assert that an object is magnetic, he notes, is to commit to an infinite number of conditional statements, none of which individually *entail* the property, but which collectively signal its presence. On this view, action emerges from rationality not by entailment but through dispositional convergence: it is the system's behavioral *signature*, not its definitional base (Hempel, 1961).

Already, we can discern the emergence of two distinct poles within the nascent action pipeline. The first—broadly computational—captures the agent's planning procedures: the simulation of action sequences based on available information, each weighted against the likelihood of achieving a desired end. The second—more primitive in structure—posits a probabilistic feature space from which these ends (and the desires that orient them) arise. Together, these layers map a substantial portion of the action terrain. Yet Hempel's framework, while structurally elegant, proves insufficient as a fully integrated interface. Most notably, Hempel concedes the model's inapplicability to non-teleological actions—those which do not consist in attaining a specified end-state. But this yields a critical question: is *action for its own sake* still action?

Moreover, Hempel's account presumes rather than explains the genesis of desire. Rationality is taken as a dispositional given—but this assumption leaves undefined the bridge between sensation-perception mappings and the motivational states that precede action. Without a more detailed account, it becomes unclear how—or even if—agents exert control over their own dispositions. This latter issue, which I take to be a modern reformulation of the classical problem of the will, is addressed most insightfully by Frankfurt in *Freedom of the Will and the Concept of a Person*. In the next section, I introduce the *Action Interface* prototype and systematically articulate each of its constituent layers, specifically elucidating how it solves the aforementioned problems.

The Prototype

The *Action Interface* comprises three interdependent modules, each responsible for a distinct layer in the architecture of agency. These modules may be delineated as follows:

I. Sensory–Perception Pipeline

A. Decomposition

B. Feature Map Construction

II. Planning Pipeline

A. Obtained-Desire Simulation

B. Distance Imposition (Single-Map Traversal)

C. Map Selection

D. Multi-Map Tuning

III. Volitional Framing Pipeline

Given the conceptual density of these components, I will proceed to articulate each module in turn—providing heuristic examples to anchor high-level intuitions while simultaneously unpacking their underlying structural mechanisms.

Sensory-Perception Pipeline

In *Intention*, Anscombe wrestles with the question of whether a desire is necessarily implicated in an intended action. For example, she writes that "I may simply hear a knock on the door and go downstairs to open it without experiencing any such desire." (Anscombe, 1956). The question then arises, *why* do we "answer the door"? Why do some mental states—such as examining a beautiful renaissance era painting—evoke in us such powerful emotions which do nothing to move us toward action, while other mental states—such as the recognition of a knock on the door—so readily carry us to move to their effect? I believe the empirical answer to this question is still out of our reach, but I will do my best to build a schematic which may serve as a pragmatic means of segmenting the different ways mental states may come about and produce actionable desires.

Decomposition

While the first Convolutional Neural Networks (CNNs) were mathematically constructed in the 1990s to perform tasks such as handwriting recognition (Kumar, 2021), the exponential rise in computational power—coupled with the widespread availability of digital imagery—has transformed the landscape of Computer Vision. Tasks that were once regarded as computationally intractable are now routinely solved by CNN architectures. Illustrated below is the VGG-16 model: a deep convolutional network that, having been trained on millions of labeled images, is capable of decomposing raw pixel data, extracting salient features, and correctly classifying an input image as belonging to a specific object type—in this case, "cat."

Figure 1

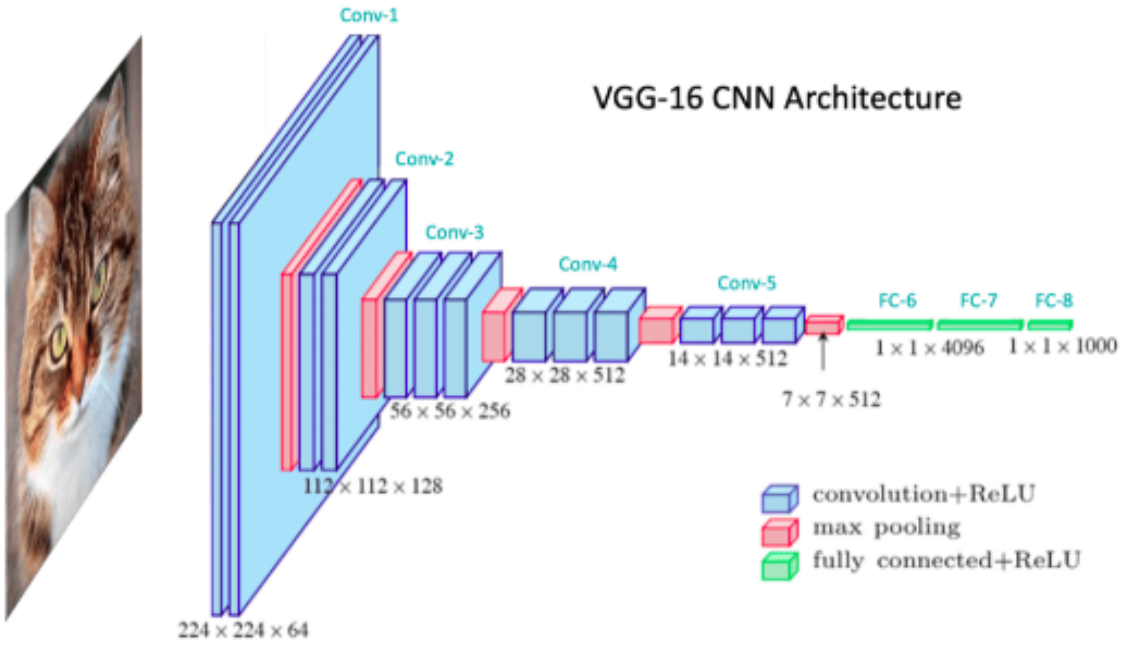
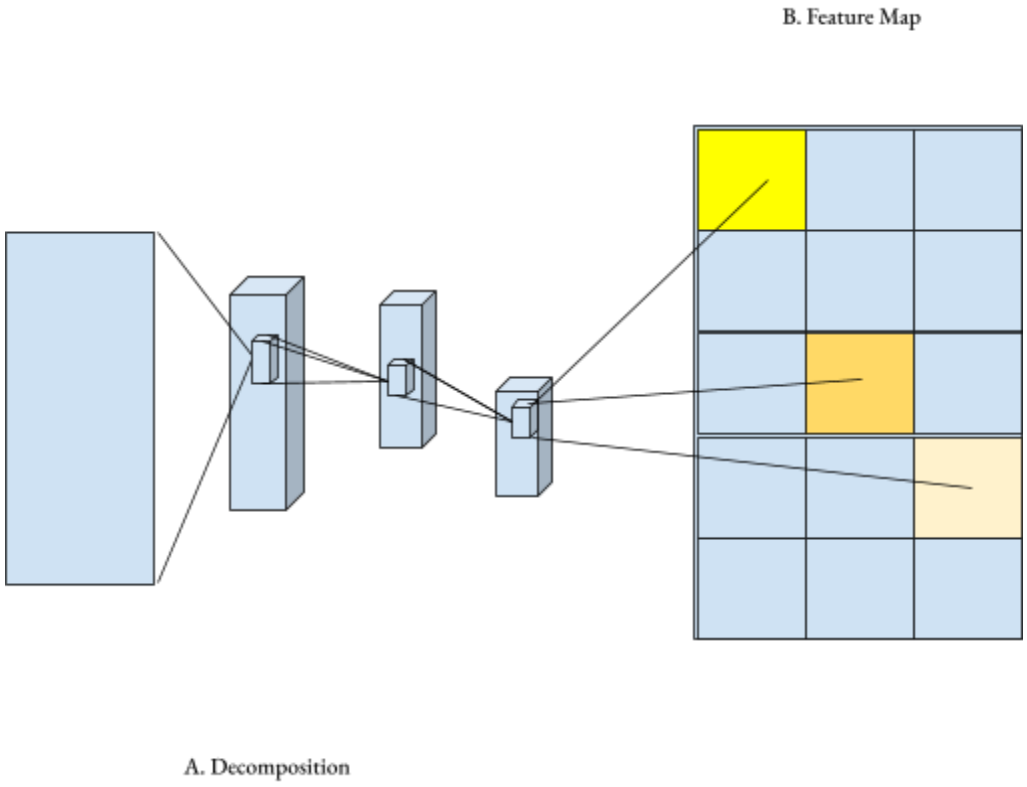


Figure found at OpenCV.com

Notably, CNNs operate through a recursive process of selective filtration. At each layer, pixel data is compressed and abstracted—irrelevant information is discarded, and high-signal features are retained. As the network progresses, dimensionality is progressively reduced, culminating in a compact one-dimensional feature vector. This vector, rich with compressed significance, is then mapped onto a classification array of matching dimensionality, yielding a final label.

It is this core structure—layered abstraction, compression, and mapping—that I now take as inspiration for the *Sensory–Perception Pipeline*. My aim is not to literalize the analogy, but to suggest a parallel: that human perception, like a convolutional system, involves the hierarchical decomposition of sensory input into increasingly abstract features, culminating in action-relevant representations.

Figure 2



In **Figure 2**, we begin with a stimulus container—a representational box indexing any form of raw sensory input (auditory, visual, proprioceptive, etc.) that requires downstream processing. Through a sequence of dimensionality-reducing transformations, this raw signal gives rise to what I will call a *mental image*: a conceptual representation of an object, scene, or pattern that is logically parsable and capable of undergoing modification. This image, however, is not yet a full mental state. Rather, a *mental state* emerges when the mental image is mapped onto a *perceptual feature map*—a dynamic space in which features are encoded as weighted frequencies. When taken in aggregate, these weighted vectors generate what we might call a *sense*: a global percept that may be affective, hallucinatory, linguistic, action-directive, or any blend thereof.

It is useful here to consider that mental states likely operate within a probabilistic range of internal coherence. That is: for any given state—assuming no introduction of novel or conflicting stimulus—the relevant feature weights tend to oscillate within a bounded Gaussian distribution. This boundedness helps explain both the *durability* and *robustness* of mental states: why they persist despite minor perturbations, and why the perceptual "feel" of a state remains relatively stable even as the environment shifts.

At this stage, it is important to qualify the scope of the model. I do *not* mean to imply that the pipeline described here corresponds directly to any particular anatomical structure or neural region. The model is, by design, an abstraction—meant to organize and clarify the logic of sensory-conceptual transformation rather than mirror the brain's architecture. Ideally, emerging neuroscientific findings will not substantially conflict with the model. But should a particular variable be shown to misrepresent sequence or function, it would be reasonable to revise the structure accordingly.

What this model aims to offer—particularly within the *Sensory–Perception Pipeline*—is a conceptual instrument for segmenting mental states: specifically, those which give rise to desire, those which do not, and those which remain inertial yet actionable. *Decomposition*, then, can be construed as a method of reliably encoding mental states from high-dimensional, context-rich inputs. It is not an empirical claim but a logical scaffolding—intended to make tractable the terrain between sensation and volition.

Feature Map

"A reason rationalizes an action only if it leads us to see something the agent saw, or thought he saw, in his action—some feature, consequence, or aspect of the action the agent wanted, desired, prized, held dear, thought dutiful, beneficial, obligatory, or agreeable" (Davidson, 1963). I take Davidson's formulation as pointing toward a core desideratum of the *Action Interface*: the ability to identify those weighted features embedded within a mental state that give rise to an *actionable desire*. For brevity, I refer to these sub-features as

Motivation Units, or μ. Each unit is indexed by a weight ω, such that the operative motivational profile may be represented as:

$$\sum_{i=1}^{n} \mu_i \omega_i$$

This weighted summation does not purport to model the mechanism of movement per se—only to identify the internal constituents which plausibly exert causal force within the mental state itself. The exact process by which a set of μ-values crosses the threshold into motor execution remains elusive, and I do not attempt to resolve it here. As previously noted, mental states are multidimensional and interdependent, implicating a wide array of perceptual, affective, and procedural subsystems. It is sufficient, for present purposes, to posit that certain feature-weight configurations possess *movement-prompting potential*, and that these configurations likely operate beneath the level of conscious access.

This claim is bolstered by empirical evidence. In a now-famous set of neurophysiological experiments, subjects were asked to record the moment at which they consciously decided to initiate a simple movement (e.g., lifting a finger). In every case, a measurable *readiness potential* was detected in the brain—hundreds of milliseconds prior to both the movement and the subject's conscious awareness of deciding to move. This anticipatory signal was consistent across trials and temporally predictive of the movement itself. As Nørretranders (1999) argues, the presence of such a signal appears to indicate that volitional action—at least in its microphysical form—is preceded by deterministic processes which unfold below the horizon of introspective access.

Planning Pipeline

At some point, the abstract indices of a hypothetical *Feature Map* must be concretized—translated into a form accessible to rational reflection and capable of guiding behavior. I hear music and feel the urge to dance. I see an advertisement for a new restaurant and

become aware of the hunger I had not explicitly noticed. I am hurt by a friend and experience a swelling impulse toward revenge. In each case, desire is not conjured ex nihilo but arises from pre-processed perceptual data—data that has already instantiated a mental state of sufficient salience to suggest movement. And yet the presence of actionable desire prompts further questions: *Will I move? What would stop me?*

Perhaps I'm in public, and embarrassment constraints my urge to dance. Perhaps I'm in a meeting, and professional obligation overrides my hunger. Perhaps my commitment to grace intervenes—quietly superseding the desire to retaliate. These tensions reveal a second-order dynamic: that desires do not merely activate behavior, but compete, collide, and in some cases, cancel. Nørretranders (1999) has speculated that *consciousness* itself may function as a kind of "veto machine"—an apparatus that allows us to *suppress* but not necessarily *initiate* action. But even this notion requires scrutiny. Would not the act of vetoing imply a reason in its own right—a second-order desire not to act?

These questions point us toward the *Planning Pipeline*, which offers a formal structure for understanding how desires—once activated—are weighted, modulated, and either executed or deferred. It is here that the logic of inhibition, redirection, and commitment finds its operational form.

Obtained-Desire Simulation

The simplest means of weighting competing desires is to simulate their satisfaction under the assumption that each has already been obtained. A mouse, presented with several varieties of cheese, samples each and discovers a preference for Cheddar. While it might not reject a nibble of Gouda or Colby, if effort were held constant and access unconstrained, it would choose Cheddar before the others. The simulation, in this case, is embodied through experience—but the logic is more general: preference is revealed not only through action, but through *counterfactual acquisition*.

Now consider a person who, feeling hungry, entertains several candidate actions: eating a burger and fries, selecting a salad, or grabbing a handful of nuts. Each potential outcome is projected through a brief, often unconscious simulation—weighted in light of various present constraints. These constraints are themselves shaped by the *feature map*: contextual variables, bodily states, social expectations, and affective tones that delimit what "satiating" might mean in this moment. Crucially, we assume that the desire is *actionable*—that it refers to an end-state which can, in principle, be obtained. People may ruminate endlessly on fantasies or ideals they neither can nor would act upon; but such simulations remain inert unless bridged by a plausible trajectory to attainment.

It is not necessary—nor even likely—that these simulations reach the level of conscious deliberation. Nor must the agent wait for the next module of the pipeline before assembling a heuristic preference structure. On the contrary, the *Obtained-Desire Simulation* phase is likely diffuse, recursive, and fast: generating a loose preference set through weighted counterfactuals before any movement occurs. What matters is that, prior to action, the agent conceives—if only dimly—of a way to bring about the object of desire. Without such a trajectory, no volitional bridge can be formed between wanting and doing.

Distance Imposition (Single-Map Traversal

Unfortunately—or perhaps fortunately—most of our desires are not immediately obtainable. I must walk to the café to get my coffee; I need to update my software before I can launch the game. These conditional dependencies complicate the assessment of moment-to-moment preferences. Is the trade-off of walking to the café justified by the higher quality beverage, or should I settle for what I can brew at home? Returning to our earlier example, the mouse now finds itself sufficiently motivated to seek the Cheddar. But we have placed it in a maze. It traverses blindly, encountering multiple dead ends, recalibrating its route until it finally arrives at the goal.

Unlike the mouse, we are spared the inefficiency of literal trial-and-error. Instead, we simulate. Presented with a prospective goal, we construct an internal maze—a plan space populated by probabilistic projections. From this simulation, we extract key heuristics: How long will the goal take to reach? How complex are the obstacles? How likely is success? And crucially: *Are there alternative routes?* This last consideration marks the entry point into a new form of evaluative structure—what I term *Map Preference*.

But if every possible map required full traversal before comparison, the process would become computationally intractable. It would be inefficient, even absurd, to run Maze A to completion before initiating Maze B, and so on. One might imagine compressing all conditions into a single, hyper-complex structure—an enormous maze with shifting gates and embedded contingencies. But I take a different approach. Rather than fully traversing individual maps in serial, I posit a lightweight mechanism for evaluating *many* possible paths in parallel. This gives rise to the next stage in the Planning Pipeline: *Map Selection*—the estimation of obstacle-weighted proximity to goal-states across a set of candidate maps.

Map Selection

Consider the mobile game *Temple Run*, in which the player must navigate a procedurally generated maze—leaping over gaps, avoiding obstacles, and collecting rewards. Prior to gameplay, one can scroll through a set of available maps, each accompanied by basic metadata: difficulty level, potential reward, visual theme. After previewing a dozen options, the player selects one and initiates the run. It is this kind of selective act—scaled and abstracted—that I refer to as *Map Selection*.

The analogy holds at the structural level. As Hempel rightly observes in *Rational Action*, there are often many rational ways of achieving a given end-state (Hempel, 1963). Map Selection, then, is the process by which these alternative strategies are preliminarily surveyed and heuristically scored. Importantly, the resolution of these maps is low: they

represent coarse estimates rather than detailed plans. This low resolution enables rapid scanning across multiple trajectories, each of which is appended to a temporary preference set and passed back to the Planning Pipeline—either for further elaboration or immediate commitment. In this way, Map Selection functions not only as a filter but as a staging ground for volitional crystallization.

Linguistic Aside

At this juncture, I wish to draw a linguistic parallel. In *The Logic of Intention Reports*, Grano attempts to formalize the semantics of the verb "to intend." He notes that "intend" shares features with both "desire" and "believe," but crucially includes an additional component absent in either: the *RESP relation*. Grano formalizes this as follows:

$$[a \text{ intends } p]^w = 1 \text{ iff } RESP(a,p) \in max[\text{Effective-Preference}(a,w)]$$

That is: an agent *a* intends *p* in world *w* if "*a* stands in the RESP relation to *p*" constitutes the highest-ranked element in a set of effective preferences. The RESP relation, as Grano defines it, encodes a disposition to *bring about* a particular proposition. Thus, intention is not merely desire with direction—it is desire under commitment, filtered through a feasibility constraint and elevated within a preference hierarchy (Grano, 2017).

Viewed through this lens, the Planning Pipeline begins to map cleanly onto the modal terrain of propositional attitudes. The *Obtained-Desire Simulation* stage approximates "to desire": an imagining of what it would be like to remove the distance between the self and what one wants. The *Single-Map Traversal* & *Map Selection* stage, by contrast, aligns with "to intend"—the elevation of a particular plan to the top of the agent's volitional stack. What remains is "to believe." Though belief is polysemous, one plausible reading is this: belief, in this context, refers to *the reason to intend*—a confidence, however

tentative, that the selected map represents reality sufficiently well such that acting upon it will yield the intended result.*

It is also crucial to distinguish between the expressions *"to intend"* and *"to have done intentionally."* At the linguistic level, this marks a shift in tense—from present to past—but the shift encodes more than just time. *To intend* implies the presence of a particular plan: modular, provisional, and structured for future execution. It is forward-facing, probabilistic, and incomplete. In contrast, *to have done intentionally* offers a retrospective framing of action—it infers that a plan *was* present, and that it was successfully implemented. The former concerns the presence of a volitional structure aimed at bringing about a state of affairs; the latter reconstructs that structure post hoc, mapping it back onto the action already taken.**

**Another way of parsing this could harken back to Davidson's definition of belief, which he describes as "believing (or knowing, perceiving, noticing, remembering) that his action is of af that kind"—where 'kind' in this case refers to a certain pro-attitude an agent stands in relation to with a particular action. However, my account differs in its level of specificity: that what is typed about the action is its planned architecture, and one's belief is precisely the degree of confidence with which one employs said plan.*

***A brief aside is warranted here. The phrase "to have done intentionally" often carries a teleological undertone—as though intentionality required that the action in question be performed in order to obtain some state beyond itself. I will return to this issue in the "Solution to Action" section, but for the moment, I ask the reader to suspend this assumption. It is entirely plausible—both linguistically and philosophically—that certain intentional actions are self-contained, needing no extrinsic goal to justify their classification as such.*

This temporal distinction invites a useful computational parallel: the logic of *process queues* as used in modern operating systems. Without delving too deeply into implementation, the basic task of an operating system is to interface between user-level instructions and low-level hardware execution—doing so in a manner that is both efficient and secure. To achieve this, the system allocates regions of memory (RAM) where discrete processes are temporarily housed and executed, mediating between the internal code of each process and the scheduling policies that determine when—and in what order—those processes are switched on.

Suppose now that a highest-preference map has been selected. To *intend* this map is to elevate it into an additional queue—an even higher-order structure in which maps are hierarchically organized and temporally spaced. The mere act of planning does not guarantee execution; to intend is to submit the plan for processing, subject to the arbitration of competing plans, shifting priorities, and contextual constraints.

Multi-Map Tuning

When I refer to *maps*—whether construed as plans, processes, or intentions—as being held in a *hierarchy*, I mean this in two distinct senses. The first, as Michael Thomson outlines in *Naive Action Theory*, concerns the *nested structure* of teleological descriptions. Actions often contain within them sub-actions, each interpretable at a different level of granularity. For instance: "I am cracking an egg in order to make an omelette" (Thomson, 2008). One might ask whether it is appropriate to say "I'm cracking an egg" or "I'm making an omelette"—or both. And if both are true, are we speaking of a single plan, a single action, or of distinct events enumerated separately? This is a foundational question in action individuation—and one I set aside for the present moment.

Instead, I wish to focus on a second dimension of hierarchy—one which operates not across sub-token decompositions of a single plan, but across *multiple candidate plans*, each (1) defined by a different level of abstraction, (2) oriented toward distinct desires, and

(3) converging or conflicting along some shared temporal or conceptual axis. Take, for example, the case of a married man who encounters an attractive stranger. He may experience a desire to initiate conversation, to pursue a romantic connection, even to engage in physical contact. And yet, this set of emergent plans may be vetoed—not because of higher-order preferences operating on the same evaluative plane, but because of an ongoing moral commitment that belongs to an altogether *different evaluative frame*. The conflict here is not between rival strategies for the same goal, but between parallel planning processes, one of which is suspended due to norm-driven incompatibility.

Hempel references this when he notes that strategies aimed at achieving a total objective are constrained by a set of operative norms (Hempel, 1963). But crucially, such norms must be processed at a sufficiently abstract level to modulate or suppress lower-level processes without requiring their re-calculation from within. In this way, hierarchical tuning operates not just laterally across plans, but vertically across *domains of reason*.

Of course, not all interactions among plan states are antagonistic. There are also cases of synergy. Drinking coffee may satisfy a sensory desire, while simultaneously enhancing cognitive performance—thus aligning with a higher-level plan aimed at productivity. These nested or convergent outcomes require a mechanism capable of arbitrating between competing and cooperating variables distributed across time and abstraction.

This brings us to the final submodule of the Planning Pipeline: *Multi-Variable Tuning*. Here, inter-plan dynamics are evaluated holistically, such that dimensional interactions—affective, ethical, cognitive, logistical—are tuned against one another. The result is an enriched context space, capable of refining or suppressing emergent processes introduced earlier in the pipeline. It is here that short-range desires encounter long-range commitments, and where local preference structures are modulated by enduring volitional scaffolds.

Volitional Framing Pipeline

I want to take seriously, for a moment, the question of where—within the architecture just described—we might locate something akin to *the will*. Despite the layered complexity of the system, from low-level feature detection to high-level abstraction and planning, there appears to be no definitive module where an agent can simply *will* an action into being. Plans may be selected, desires scored, processes queued or terminated—but all within a structure that is, if not strictly deterministic, at least computationally enclosed. Even within *Multi-Variable Tuning*—our most abstract evaluative layer—the emergence of high-level ideals does not necessarily afford the agent an outlet to *will those ideals into motion*.

In *Freedom of the Will and the Concept of a Person*, Frankfurt makes a crucial distinction: persons are marked not merely by first-order desires, but by *second-order volitions*—that is, the capacity to *want a certain desire to be one's will* (Frankfurt, 2018). His paradigmatic example is the addict: he experiences a first-order desire to consume his drug of choice, while simultaneously possessing a second-order volition *not* to act on that desire. He is not neutral to his own wanting. He can witness the subroutines of his planning pipeline unfold—watch the desire rise, escalate into a plan, and initiate execution—all while experiencing a top-level volitional resistance that, tragically, fails to interrupt the sequence.

It is my belief that, even if we were to fully formalize the Planning Pipeline in a way that tracks with psychological ground truth, the structure itself would remain *inert* to the problem of volitional override. That is: the pipeline, taken alone, cannot explain genuine behavioral change. If an addict overcomes his compulsion in one instance, it is no guarantee that he will do so again. The system remains vulnerable. And so, we take up auxiliary strategies—*escape actions*—designed to avoid triggers, minimize exposure, or temporarily reroute desire. These interventions are often framed as success stories. But even then, they betray a fragile architecture. The addict who attends a retreat, restructures his habits, and genuinely sheds the desire for the drug may still relapse upon returning to the very neighborhood where his prior circuits were first laid down.

This return—this reactivation of dormant paths—reveals something essential. *Second-order volitions*, even if formulated in the highest strata of the planning pipeline, must reach downward to *reconfigure the Sensory–Perception Map itself* if they are to generate sustained behavioral change. That is: they must rewrite the conditions under which lower-level desires arise. This, I believe, is what we mean by a *change in perspective*—and it is likely the most difficult thing a person can will. It is not enough to revise a plan. One must *restructure the interpretive interface* through which stimuli are parsed, categorized, and assigned salience.

If you hate Sarah today, and every time you see her a hardness rises in your chest, how do you change? It cannot simply be a post-hoc update in the planning cycle—some intellectual override that retrofits the feeling into a new frame. Instead, the very *experience of seeing Sarah* must be re-interpreted at the perceptual level. Her face, her voice, the inflection of her laugh—all must be remapped. *That* is will—not the mere suppression of action, but the deep restructuring of the conditions under which action becomes possible.

The Final Theorem & Means of Testing

So, after all that, what *is* action? I've supplied a comprehensive way of thinking about how desires might arise, mature into plans, interact with other plans in a chaotic space, and abstract into forms which are capable of restructuring the entire model. However, I haven't yet formally written my answer to the Problem of Action. In this section, I will do just that—beginning with my solution in plain terms, then providing a conceptual test capable of sorting actions by intentionality.

The Solution to Action

An action is an instance of the execution of a pattern of movements which are caused by an actionable desire—which is itself motivated by certain features of a particular mental state—that is defined within a certain scope at the initial time of the action (t_0). This scope includes the following:

1. The set of features which motivated the actionable desire
2. The highest level of analysis under which the action can truthfully be described
3. A set of nested descriptions including movements and/or token-actions which the action encompasses

The first clause refers to a motivated reasoning which synthesizes Anscombe's mental cause with Davidson's desire-localized causal account. Betty heard music, and she began to dance. Decomposing this sentence into its scope will include

1. The music motivated the desire to dance
2. Dancing
3. The range of movements characterized by the verb "to dance"

We might say that, in Davidsonian terms, the music incited in her a certain recognition of a desire of which she held a pro-attitude toward, and her movements were the expression of a belief that she was performing an action of that kind. But what of Frankfurt's criticism of this line of reasoning? Let us consider one of Frankfurt's prime challenges: the case of basic causal deviation. In an example he provides, "a man at a party intends to spill what is in his glass because he wants to signal his confederates to begin a robbery and he believes, in virtue of their prearrangements, that spilling what is in his glass will accomplish that; but all this leads the man to be very anxious, his anxiety makes his hand tremble, and so his glass spills." (Frankfurt, 1978).

The key in reconciling Davidson with this example is to understand when (a) antecedent causes must be updated to account for new information, and (b) to identify those new causes as inadequate for action. In order to meet these conditions, allow us to first consider the scope of the initial ine-spilling case

1. The intention to incite the robbery motivated the desire to spill the wine
2. Inciting the robbery
3. The token action of spilling the wine

However, at the moment prior to spilling, if we were to assess the scope, we would find that key descriptions have changed. If we were to attempt to restructure the scope, we'd find

1. The man's anxiety caused him to spill the wine
2. Spill the wine
3. The range of movements characterized by "spilling the wine"

There are two distinct things to point out here. Firstly, we lose the teleological factor. The level of analysis is no longer characterizable under the higher level description of "intending to incite the robbery". Furthermore, the causal agent "anxiety" did not trigger an actionable desire. Not only can we say that the man did not intend to incite the robbery by his action, but I'd argue that this was no action at all, for precisely the same reason that it *would* have been an action if there was no disturbance, and the man indeed moved his body to the effect of the conditions set in the original scope.

Therefore, an action is terminated (and/or must be re-framed) when an upstream descriptor previously defined by the scope changes. For example, if I feel parched and want to drink water, I may then enumerate the token-actions I take by walking downstairs, pulling out a glass from the cupboard, dispensing water into it from the faucet, then taking a sip. However, if instead, while walking downstairs, I realize I am no longer parched, but I want to water my plants and continue to take out a glass and dispense the water into it, those actions—while still actions—proceed from different antecedents, and so require an updating in scope. We could even granularize this example further and say, even though my primary reason for getting the water was to water the plants, I decided to take a sip on my way to the deck. The solution to this problem would be either (a) the scope actually included both features as motivation for my action, such that my thirst and intention to water my plants motivated my desire to fetch water whereupon I proceeded to drink and then water my plants, or (b) we fork the scopes separately and simply provide a modal account of each action.

Turning now from motivation to intention, I want to demonstrate how the identification of the scope of an action can help us correctly identify which descriptions of intentionality hold. Consider the following sentences

A. Richard was *angry at Sharon*.
B. Richard was so angry at Sharon, *he ripped up some paper*.
C. Richard was so angry at Sharon, he ripped of some paper *of hers*.S
D. Richard was so angry at Sharon, he ripped up her *appointment slip*.

As a preliminary note, we should highlight that (D→C), but ¬(C→D). This is true for each subsequent traversal down the description list. Therefore, we can frame intentionality by reference to the scope at the time of action. Suppose we consider "C" to be the true description. We'd say

1. Richard was angry at Sharon motivated the desire to rip up some paper of hers
2. Sharon's paper
3. Rip up some paper

It then becomes false to claim "D" since "appointment slip" is outside of the scope of the action description.

It is important to note that intention has two usages. The first denotes a binary "was/was not intentional", while the second is a graded value subject to adjectives like "more" as in "more intentional". In my view, these distinctions are actually just short-hand for indicating the level of specificity with which an intention was executed. While it might be true that "Richard intentionally ripped up some paper", we might say the action was more intentional if in fact the correct description was "Richard ripped up Sharon's appointment slip", since the latter description indicates a more specific plan. In the next section, I will provide a conceptual test—predicated on the theorem—which can be used to differentiate actions which are intentional from actions which are not.

The Counterfactual Substitution Test

In *Intention*, Anscombe ruminates on what distinguishes actions which are intentional from those which are not. She asks us to consider the question " 'Why did you knock the cup off the table?' [which is] answered by 'I thought I saw a face at the window and it made me jump.' " (Anscombe, 1956). I wish to answer this question in a way that also solves the issue of consequential deviance and builds a framework for parsing intentionality chains.

When we are saying of an action that it is intentional, what we are really saying is that, at the time of acting, the actor had enumerated certain key properties (which I will refer to as "targets") that, should they have been absent, the actor would have seen no reason to act in that particular way, and whose presence signals (a) a necessary, causal factor in the agent's motivated desire to act, and/or (b) an object which the actor wanted to effect in a specific way.

Returning to Anscombe's example, allow us to presume that we wish to determine whether or not "knocking the cup" was intentional. What we should do is then label "the cup" as the target in question, one whose presence we presume was a necessary, causal factor in the agent's execution of her actionable desire to knock it over. Then we consider a series of possible worlds wherein we piecewise substitute "the cup" for other objects. Suppose, instead of a cup, there was a block of wood with the same dimensions, or a standing remote, or some other device whose properties differ from that of the cup. Then we only need to ask, in these possible worlds, would the agent still act? From this question we can consider three possible answers:

1. The agent would not act in any other world other than the one in which that particular cup was present.
2. The agent would act in *some* of the other worlds in which the cup was substituted.
3. The agent would act in *all* (or a sufficient number of other worlds such as any distinguishing properties of the cup are effectively nullified) of the other worlds.

In Scenario 1, the object in question is *specifically implicated* in the causal chain leading from desire to action. In Scenario 2, there exists some *typed description* of which the target in the evaluative world is a member, and it is this particular type which motivates the causal transformation from desire to action. In Scenario 3, we identify the target as *non-causal* in motivating action, and therefore that particular target is *non-intentional*.

Applying this logic to the cup example, we would find that the actor would have executed the knocking motion even if the cup were substituted with a large range of other objects, precisely because the cup was not causal in motivating the action.

Let us also consider another famous example, one of consequential causal deviance. Suppose Jim is holding a rifle with the intention to shoot and kill Steve. In line with his plan, he raises his rifle and pulls the trigger, but missing by a large distance, he stirs up a pack of bulls which proceed to stampede Steve. The question arises: did Jim *intentionally* kill Steve? Applying the Counterfactual Substitution Test to this case, we'd find:

I. Jim pulls the trigger (shoots) in a world where Steve is present
II. Jim doesn't pull the trigger in a world where Steve is substitute
III. Jim pulls the trigger in a world where the bulls are present
IV. Jim pulls the trigger in a world where the bulls are substituted

What this result tells us is (1) Steve is an intentional target, but (2) the bulls are not, and since the bulls are implicated in the causal chain of events which lead to validating the proposition "Jim kills Steve", it scans false for intention. This is not to say we can't reason about a degree of culpability, after all Jim *did* intentionally *attempt* to kill Steve (and would have succeeded should his aim have been better), but such cases of probability require us to draw a line with regard to intentionality and the level of specificity at which we can make true claims about it.

This is not to say we can't find intentionality across causal chains which extend beyond mere action:effect. Take, for example, Chase, a high school student who desires to hurt an enemy, Richard. To do so, he plans to tell Richard's girlfriend, Carol, a lie—and he

believes that by virtue of this rumor, it will cause an argument to ensue between Carol and Richard such that his primary goal is achieved. In a case such as this, we would find that Richard is specifically causal in motivating Chase's action, and Carol is at least causally implicated at the level of type (perhaps another close friend would have sufficed). Therefore, the definition of intentionality is merely to apprehend certain salient targets at the time of action, and that those targets (upon completion of the event) contain the complete set of actual, causal factors which led to the original desired end state.

In summation, an action can be viewed from the lens of a temporal process: one which is motivated by certain features and leads to the execution of an actionable desire. The scope of the action is defined at the moment of action by these antecedent causes, and must be terminated or re-evaluated if the scope is violated during the action event. We can think of intentionality as a set of targets which were causally implicated in motivating the desire to act and which constitute the necessary factors which brought about the desired end state of the action event.

Conclusion

The Problem of Action constitutes an ongoing philosophical discussion about the nature of action and how to differentiate it from mere happenings. The terrain includes theories of causation, guidance, knowledge, teleological realism, and rationality. Early in this paper I identify the need for a means of communicating about the problem which can make sense of these disparate theories. I believed Hempel's empirical hypothesis was a useful heuristic to build upon, serving as the basis for an Action Interface that could connect empirical, rational, and propositional endpoints. After enumerating the features of the interface's three main branches, I offer my own solution to the Problem of Action: one which extends Davidson's theory of causal antecedents as arising from certain features of emergent mental states (which themselves are linked to more primitive dispositional attitudes), and attempt to reconcile this view with Frankfurt's critique by defining a scope at the outset of action that—when broken—would necessitate a re-analysis of the

causal markers driving the agent's new pattern of movement. Finally, I offer a formal way of defining intentionality: that an action is only intentional under a given description if all necessary factors for achieving the end state are causally motivating of the desire which moved the agent to act.

References:

Ackerman, J. M., Huang, J. Y., & Bargh, J. A. (2012). Evolutionary perspectives on social cognition. *The handbook of social cognition*, 451-473.

Anscombe, G. E. M. (1956). Intention. *Proceedings of the Aristotelian Society*, *57*, 321–332. http://www.jstor.org/stable/4544583

Davidson, D. (1963). Actions, Reasons, and Causes. *The Journal of Philosophy*, *60*(23), 685–700. https://doi.org/10.2307/2023177

Frankfurt, H. G. (1978). The problem of action. *American philosophical quarterly*, *15*(2), 157-162.

Frankfurt, H. (2018). Freedom of the Will and the Concept of a Person. In *Agency and Responsibility* (pp. 77-91). Routledge.

Grano, T. (2017). The logic of intention reports. *Journal of Semantics*, *34*(4), 587-632. https://shorturl.at/NmHZx

Hempel, C. G. (1961, January). Rational action. In *Proceedings and Addresses of the American Philosophical Association* (Vol. 35, pp. 5-23). American Philosophical Association. https://www.jstor.org/stable/3129344

Kromydas, B. (2024, October 25). *Convolutional Neural Network: A complete guide*. LearnOpenCV. https://learnopencv.com/understanding-convolutional-neural-networks-cnn/

Kumar, B. (2021, August 31). *Convolutional Neural Networks: A brief history of their evolution*. Medium. https://medium.com/appyhigh-technology-blog/convolutional-neural-networks-a-brief-history-of-their-evolution-ee3405568597#:~:text=The%20name%20convolutional%20neural%20networks,the%20handwritten%20digit%20recognition%20task.

Shepherd, J. (2021). *The shape of agency: Control, action, skill, knowledge* Oxford University Press.

Thompson, M. (2008). Naïve action theory. *Life and Action*, *2010*, 85-148.

Tor Nørretranders. (1999). *The user illusion : cutting consciousness down to size*. Penguin.