OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

A Formalization and Defense of Basic Prudential Hedonism

Gaocheng Zeng
zenggaocheng@gmail.com

ABSTRACT

This paper aims to formalize Basic Prudential Hedonism (BPH)-which holds that an agent's prudential obligation lies in maximizing net pleasure and constructs a tailored instrumental obligation framework for ignorant, learning hedonistic agents. It first clarifies BPH's philosophical foundations (e.g., prudence-morality separation, hedonic welfare) and defends its three core theses (phenomenological, calculation, evaluative) against objections via Nelson's arguments (|Nelson, 2020, p. 15]). Then, it sketches a modification for Yan and He's (Yan and He 2025) causaldeontic model of instrumental obligation to include hedonic value quantification, belief weighting, probabilistic interventions, and learning mechanisms (Bayesian updates). Finally, it validates the "Ought Implies Can" principle for BPH by distinguishing prudential from moral obligations, addressing King's (King, 2014, p. 316]) challenge. This work provides both theoretical clarity for BPH and a practical tool for analyzing prudential choices.

1 INTRODUCTION

Prudence - the capacity to govern one's actions for the sake of one's own well-being-has long been a central concept in ethical thought, though its relationship to morality has shifted over time. Aristotle originally framed prudence as intertwined with moral virtue, arguing neither could exist without the other (Aristotle, 2011, 1144b30-32]); later, Kant separated the two, casting prudence as conditional "imperatives" tied to individual happiness, distinct from the universal, unyielding demands of moral duty (Kant, 2017, Section II]). Within this tradition, hedonism has emerged as a key foundation for prudential theory. Basic Prudential Hedonism (BPH) holds that an agent's well-being depends solely on net pleasure (pleasure minus pain), and prudential obligations should center on maximizing this value. Yet BPH faces critical gaps: First, it lacks systematic formalization to anchor its core claims in actionable logic; Second, existing models of instrumental obligation (which guide how to act to achieve goals) fail to account for real-world hedonistic agents-who often lack certainty about causal links (like whether a behavior will yield pleasure) and learn from experience; Third, there is no clear framework to translate BPH's "maximize net pleasure" goal into concrete, step-by-step decision-making. For instance, a person deciding between staying up late watching a show (immediate, shortterm pleasure) and going to bed early (mild short-term discomfort but long-term energy and better mood the next day) often struggles to judge which choice truly maximizes net pleasure. BPH's broad goal of "maximizing net pleasure" gives no guidance on how to compare the intensity (the show's excitement vs. morning tiredness) or duration (the show's two hours vs. a full day of energy) of these conflicting outcomes-let alone account for uncertainty (e.g., Will the show be as enjoyable as expected? Will I really feel tired tomorrow?).

This backdrop shapes the core research context of this paper: to address these gaps, we aim to construct an intelligent agent that takes BPH as its fundamental goal (maximizing net pleasure) and instrumental obligation as its practical means (determining which actions will reliably achieve that goal). This exploration carries distinct value for the AI field: current AI systems often focus on external, task-specific objectives (e.g., completing a task, optimizing efficiency) but lack a rigorous framework for modeling "agent-centric welfare" (or the welfare of the humans they serve). Integrating BPH and instrumental obligation fills this gap, offering a theoretical basis for designing AI with human-like prudential reasoning-such as personal health management AI that prioritizes long-term net well-being, or adaptive life assistants that balance immediate comfort and future pleasure.

This paper addresses these gaps with three core goals: First, it formalizes BPH by clearly defining its three foundational theses: the phenomenological thesis (all pleasures and pains share a common defining property), the calculation thesis (hedonic value can be quantified by intensity and duration), and the evaluative thesis (welfare equals net hedonic value). It also defends these theses against key objections, including claims that pleasures lack a shared property and that non-hedonic factors (like real friendships or life trajectory) inherently shape welfare; Second, it modifies an existing causal-deontic model of instrumental obligation to fit BPH's needs, adding tools to quantify hedonic value, weight causal beliefs by certainty, calculate expected pleasure under uncertainty, and update beliefs and values as agents learn from experience; Third, it validates the "Ought Implies Can" principle for BPH, resolving challenges by emphasizing that prudential obligations are tied to an agent's actual abilities - including knowledge, physical capacity, and access to opportunities - and differ from moral obligations in their focus on individual self-interest rather than universal rules.

The paper proceeds in six sections. Section 2 clarifies key concepts that underpin BPH , including the distinction between prudence and morality, the nature of hedonism (descriptive vs. normative), the link between pleasure and welfare, and the structure of beliefs, desires, ability, and obligation; Section 3 formalizes BPH's core theses and defends them against both non-evaluative objections (about the nature of pleasure) and evaluative objections (about what contributes to welfare); Section 4 introduces the original causal-deontic model of instrumental obligation, using a weight-loss case to illustrate its components: causal models, interventions to test action outcomes, and priority structures to rank goals; Section 5 extends this model to BPH, integrating hedonic value quantification and learning mechanisms; Section 6 examines the "Ought Implies Can" principle in the context of BPH, addressing challenges and linking prudential obligation to agent ability; Section 7 summarizes the formalized BPH framework and its practical implications for analyzing prudential choices-including the design of the BPH-instrumental obligation agent outlined here.

2 PHILOSOPHICAL PRELIMINARIES

2.1 Prudence vs. Morality

Through time, morality has been analyzed as virtue, duty, mean to good consequence, sentiment, social contract, product from social context, etc. Regardless of its analysis, however, morality has always involved the categorization of actions into those which are right and those which are wrong ([Long and

Sedley, 1987, pp. 366-367]). While studying morality provides the "what" of right or wrong, prudence, as the ability to govern and discipline oneself by the use of reason (Merriam-Webster 2025), and thereby promote one's own well-being, addresses the "how". Through history, there has been a trend that philosophers detach prudence from morality.

2.2 Hedonism

Hedonism holds that what is fundamentally good is pleasure (and the absence of pain), and happiness consists in a state where this pleasure outweighs pain. Richard Crisp argues that pleasure is the only thing we consider good, and pain is the only thing we consider bad ([Crisp, 2006a, pp. 619-645]); hedonism involves both descriptive and normative perspectives. As for the descriptive perspective, psychological hedonism describes our mental state as exclusively driven by the attainment of pleasure and the avoidance of pain (usually, their own) ([Mandle and Reidy, 2014, p. 336]); As for the normative perspective, Axiological hedonism talks about value judgments-that pleasure is the only thing intrinsically good; prudential hedonism, on the other hand, offers a norm that the well-being we pursue should be pleasure over pain (as far as self-interest is considered). There are various kinds of prudential hedonism, and we argue that basic prudential hedonism is the best theory in Section 3.

2.3 Welfare

In hedonism, pleasure is the only thing that is good for anyone, and pain the only bad, together influencing welfare (well-being) (Crisp 2006a]); for instance, someone can get pleasure from drinking a cool glass of lemonade or completing her first reading of Jane Austen's Pride and Prejudice ([Crisp, 2006a, p. 632]). There is a debate regarding whether welfare comes from the thing itself or subjective feelings, and whether welfare differs due to quality; I will propose one perspective in Section 3.

2.4 Beliefs

2.4.1 Naïve Belief/Intuitive Belief

Intuitive beliefs are a fundamental cognitive category in the mind, stored in a "belief box". The concept of the "belief box" is introduced by Schiffer (Schiffer 1981) to explain how beliefs are represented and processed in the mind. In this model, beliefs are akin to mental sentences or propositions that are "tokened" within a cognitive system. This metaphor suggests that beliefs are not isolated mental states but are part of a dynamic system
where they interact with other mental representations. The "belief box" serves as a functional placeholder for the role beliefs play in our cognitive architecture.Intuitive beliefs originate from two automatic, non-reflective processes: direct perception and spontaneous, unconscious inference; any representation stored within it is automatically treated as a "representation of actual states of affairs" ([Sperber, 1997, p. 68]). Concerning perceptual intuitive beliefs, when you see autumn leaves under a tree and form the belief "There are leaves under this tree", the origin (your visual perception of the leaves) is an automatic sensory process-you do not consciously "decide" to trust this perception or reflect on its source; the belief simply enters the belief box as a byproduct of perception ([Sperber, 1997, pp. 77-78]); concerning inferential intuitive beliefs, when you hear a doorbell ring and infer "There is someone at the door", the inference is

spontaneous and unconscious-you do not consciously retrieve a premise like "Doorbells ring when someone presses them" or reflect on how you arrived at the conclusion. The resulting belief is stored in the belief box without any conscious trace of its inferential origin ([Sperber, 1997, pp. 77-78]).

Since the origins of intuitive beliefs are already unconscious and automatic, storing them in the belief box (which treats them by default as valid) makes deliberate reflection on origin redundant, because there is no "conscious origin" to reflect on. Therefore, as for intuitive beliefs, their status as "valid data" for cognition does not depend on explicit reflection about how or why they were formed. They are immediately available as premises for practical and epistemic inferences.

Sperber (Sperber 1997) argues that without the belief box, trying to validate beliefs by reflecting on their origin would lead to a fatal logical paradox: infinite regress. Suppose, for example, you formed the belief " P " and felt the need to reflect on its origin ("Why do I believe P?") to confirm it. This would require another belief about the origin ("I believe P because I saw it"), which would then require reflection on its origin ("Why do I believe I saw it?"), and so on indefinitely. The belief box breaks this regress by grounding the validity of intuitive beliefs in their location, not in a chain of origin-based justifications.

Also, intuitive beliefs are designed (by cognitive architecture) to act as fast, ready-to-use premises for everyday reasoning and behavior, undermining the need to reflect on them. Seeing a cat approach triggers the intuitive belief "A cat is coming", which immediately activates a mouse's flight plan ([Sperber, 1997, p. 68]); seeing a friend frown triggers the intuitive belief "My friend is worried", which guides your decision to offer comfort ([Sperber, 1997, pp. 77-78]).

2.4.2 Reflective Belief

A reflective belief is a propositional belief you hold after explicitly representing, evaluating, or endorsing the belief from a higher level-i.e., you not only have the first-order content ("p") but also entertain a second-order attitude about it (e.g., "I have good reasons to believe p", or "it's justified that p") ([Sperber, 1997, pp. 71-72]); reflective beliefs and the belief box are completely isolated, making their validity not rely on locations in the belief-box, but rather external validating contexts (e.g., authority, argumentation) instead of location. A student believes "Water is H_2O " because they accept the validating context of "scientific definition" ([Sperber, 1997, p. 79]); this belief is not stored in the belief box; instead, it exists as a meta-representation (e.g., "The textbook (reliable) says water is H_2O ") stored in general memory, not the belief box.

Reflective beliefs are changeable: the change of validity of the validating context changes the reflective belief. Before 1994 (when it was unproven): The reflective belief in Fermat's Last Theorem relied on the validating context: "It is a very-well-supported conjecture" ([Sperber, 1997, p. 82]); individuals believed it not because of proof, but because of accumulated mathematical evidence and consensus. After 1994 (when Andrew Wiles proved it): The validating context shifted to "It is a theorem" ([Sperber, 1997, p. 82]); the content of the belief ("For $n > 2, x^n + y^n = z^n$ has no integer solutions") stayed the same, but the reflective belief itself changed, making the belief more credible.

Changing these contexts requires the ability to reject, modify, or sophisticate the underlying context. A student, Lisa ([Sperber, 1997, p. 75]), might initially form a reflective belief in "millions of suns" based on her teacher's authority ("The teacher said there are millions of 'suns'"). But as she learns astronomy, she replaces the "teacher authority" context with "stellar physics" (e.g., "Suns are stars, and telescopes confirm millions of stars")-refining her reflective belief into a more precise, evidence-based one.

Reflective beliefs often involve "reflective concepts"-concepts understood only via explicit theories. For example, " H_2O " can be understood only if one learns chemistry, and "prime factor" can be understood only if one learns number theory ([Sperber, 1997, p. 79]); these concepts can transform between reflective and intuitive status, which in turn changes the nature of the beliefs built around them. Here is a case when reflective belief can turn into intuitive belief: A student first learns "odd/even numbers" as a reflective concept (needing deliberate calculation to identify; Sperber, 1997, p. 80]); with practice, the concept becomes intuitive - they can spontaneously judge " 7 is odd" without reflection. The belief " 7 is odd", once a reflective belief (dependent on "teacher explanation"), becomes an intuitive belief (grounded in the belief box); By contrast, an intuitive belief can also turn into a reflective belief: A child initially has an intuitive belief in "weight" (e.g., "Heavy things fall faster"). When learning physics, they must distinguish "weight" from "mass"-a reflective concept. Their old intuitive belief is replaced by a reflective belief: "Weight is the force of gravity on mass, and falling speed depends on air resistance" ([Sperber, 1997, p. 80]).

In the context of prudential hedonism, rational agents can generate epistemological possibilities of whether their beliefs are true based on both their mind's capacity to generate spontaneous inferences from perception/experience and the meta-representational ability to evaluate representations of possibilities.

2.5 Desires

In the context of prudential hedonism, desire involves "desire" and "desire not": both refer to people's attitudes toward what they believe will give them pleasure or pain. Desire is an expectation, a belief of something not yet happening. However, in basic prudential hedonism (mentioned in Section 3), desire is an overly redundant concept, and we can avoid this concept in the course of formalizing basic prudential hedonism.

2.6 Ability

Philosophers' discussion regarding ability qualification focuses on four factors: knowledge, capabilities, access, and luck. Specifically, to perform action A, one must have the relevant knowledge of how to perform A; one must have the necessary abilities-both physical and psychological; there must be access in the sense of opportunity and non-blockage; luck
must not intervene by way of "strange" or unforeseen barriers.
Fisher posits that ability is tied to what the agent knows (or has reasonable access to know) and the internal capacities to act (Fischer 2012); Kadri Vihvelin (|Vihvelin, 2013, pp. 10 15]) distinguishes narrow ability: Adds to knowledge the physical and psychological capabilities (skills, competence, know-how). For instance, one may know how to ride a bicycle but be physically paralyzed - they then lack the narrow ability to ride; Vihvelin also defines Wide ability (or specific ability) as narrow ability plus access: the

external circumstances must permit the exercise of that capability (e.g. there must be a bicycle, it must not be locked away, etc.).

Mele (Mele, 2007, p. 49]) distinguishes between different ways in which luck can compromise ability or responsibility. For example, present luck (e.g., fluctuations in psychological states) might affect what one is able to do. If, due to bad luck, one lacks physical or mental capacity at a given moment, that limits ability.

To sum up: one is able to perform action A if and only if they have the relevant knowledge of how to perform A, the physical and psychological abilities to do so, relevant access (opportunities), and no "strange" events prevent them from performing A. When agents judge whether they have the ability, they can use the above criteria. For example, "being able to ride a bike" is analyzed as: knowing how to ride a bike, having legs and a normally functioning body and mind, a bike available, and no one threatening to harm the agent if they try to ride.

3 THE APPLICATION OF BASIC PRUDENTIAL HEDONISM

3.1 Theses

3.1.1 The Phenomenological Thesis

The phenomenological thesis describes that all pleasures share a common phenomenal property (e.g., a universal "feeling of pleasantness" that defines them as pleasures) and all pains share a distinct phenomenal property (e.g., a universal "feeling of unpleasantness" that defines them as pains) ([Nelson, 2020, p. 22]). For example: the pleasure of solving a difficult math problem (mental, intellectual) and the pleasure of eating creamy chocolate (physical, sensory) feel different in details-but BPH argues they both share the core "pleasantness" that makes them pleasures. Similarly, the pain of a stubbed toe (sharp, localized physical) and the pain of grieving a loss (dull, emotional) share the core "unpleasantness" that classifies them as pains.

3.1.2 The Calculation Thesis

The calculation thesis specifies that hedonic states (pleasures/pains) have magnitudes determined solely by their intensity (how strong the feeling is) and duration (how long the feeling lasts). These magnitudes are quantifiable in principle via cardinal numbers (e.g., a $5/10$ intensity pleasure lasting 20 minutes has a magnitude of 100) and commensurable within types-all pleasures can be compared to other pleasures, and all pains to other pains ([Nelson, 2020, p. 35]); for instance: A 10-minute burst of intense joy (intensity $= 8$) has a magnitude of 80 , while a 40 -minute feeling of mild contentment (intensity $= 2$) also has a magnitude of 80 -per BPH, these two pleasures contribute equally to a person's welfare.

3.1.3 The Evaluative Thesis

The evaluative thesis ties welfare directly to these hedonic magnitudes: a person's total welfare over a lifetime (or a segment of life) equals the sum of their pleasure magnitudes minus the sum of their pain magnitudes. This makes welfare purely quantitative - no other factors (e.g., friendship, achievement,

morality) affect welfare in itself (Nelson, 2020, p. 48]); for example: If Person X experiences 500 units of pleasure and 200 units of pain in a year, their annual welfare is 300 units; if Person Y has 400 units of pleasure and 100 units of pain, their welfare is also 300 units. Even if X has more friends or Y has more achievements, their welfare is equal because their net hedonic magnitudes are the same.

3.2 Non-evaluative Objections

There are objections claiming BPH conflicts with non-evaluative facts about pleasure/pain. Critics like Sidgwick ([Sidgwick, 1981, p. 127]), Alston (Alston, 1967, p. 344]), and Feldman (Feldman, 2006, pp. 83-84]) argue introspection reveals no common phenomenal property across pleasures/pains. For example: Alston points out that "enjoying playing tennis" (physical, active) feels drastically different from "getting satisfaction from seeing an enemy in distress" (malicious, psychological)-and both feel nothing like "the sense of well-being from a good night's sleep" (calm, passive). He claims there's no single "pleasantness" tying these together. Feldman uses sensory pleasures to make the same point: The pleasure of smelling salty ocean air (olfactory) and the pleasure of feeling warm sunlight (tactile) have "nothing phenomenologically in common"-one is a smell, the other a bodily feeling-so there's no shared property defining them as pleasures. Critics like Brentano ([Brentano, 2009, pp. 30-31]) object that magnitudes are unquantifiable. Brentano uses an example: "The pleasure I get from smoking a good cigar, multiplied by 127, equals the pleasure of listening to Beethoven's Ninth Symphony." This sounds absurd, he argues, to show hedonic magnitudes aren't quantifiable.

Nelson counters by rejecting "qualitative insight"-the assumption that introspection either reliably detects shared phenomenal properties or reveals numerical ratios ([Nelson, 2020, p. 55]); Nelson argues that pleasure/pain "permeate" other sensations, hiding their common property, describing this as "phenomenological overlap": pleasure and pain rarely exist in isolation - they are almost always intertwined with other sensory or mental states (e.g., the pleasure of a symphony overlaps with music, emotional engagement, or memory; the pain of a headache overlaps with head tension or irritability). This interweaving makes it impossible to "isolate" pure hedonic magnitude via introspection-so assigning exact ratios (e.g., 127:1) feels absurd, not because magnitudes don't exist, but because we cannot disentangle them from other states.

Moreover, Nelson remains agnostic between two sub-theories compatible with the phenomenological thesis: the Separate experience view, claiming that pleasure/pain are isolated feelings ([Bramble, 2013, p. 210]); and the Hedonic tone view, claiming that pleasure/pain are tones of other sensations ([Broad, 1930, pp. 229-230]), finding objections to both inconclusive. ([Nelson, 2020, p. 62]).

3.3 Evaluative Objections

For evaluative objections (claiming non-hedonic things have intrinsic value), Nelson uses an "undermining strategy" (Nelson, 2020, p. 70]);Some critics rely on "life comparisons":

Fletcher ([Fletcher, 2016, pp. 15-16]) and Lin ([Lin, 2016, p. 321]) compare Person A (real world: job, friends, hobbies, pleasure from these) to Person B (lifelong experience machine simulating A's pleasures, no real interactions). A and B have equal total pleasure/pain-but intuition favors A. Velleman ([Velleman,

1991, pp. 49-50]) contrasts Person A (upward trajectory: poverty → troubled youth → struggles → success → retirement) with Person B (downward trajectory: blissful childhood → precocious success → disasters → misery). A and B have equal pleasure/pain sums-but intuition favors A.

Nelson's pro-hedonist hypothesis explains these intuitions without rejecting BPH: we mistakenly assign intrinsic value to non-hedonic things (friendship, reality, life upwardness) if they are cnsistently conducive to net pleasure, not instrumentally defined (e.g., not money), and not done for external ends (e.g., not eating for nourishment) (Nelson. 2020, p. 75]);

For example, "reality" (A's life) is linked to more long-term pleasure: Real friendships provide ongoing emotional support (future pleasure), while machine friendships vanish if the machine breaks. We mistake "reality" for intrinsic value, but we're responding to its instrumental link to more pleasure. Moreover, "Life trajectory" (upward vs. downward) is linked to pleasure: Upward trajectories involve increasing pleasure (e.g., success brings more joy over time), while downward trajectories involve increasing pain. We think "upward" is better in itself, but we're reacting to the expected pleasure of an upward path.

On the other hand, some critics use examples to emphasize qualitative factors. In Roger Crisp's "Haydn and the Oyster" scenario (Crisp, 2006b, p. 112]; Nelson, 2020, p. 100]), critics argue Haydn's life is better than an oyster's infinite mild pleasure - proving "quality" matters more than quantity, which contradicts BPH's Evaluative Thesis. Nevertheless, Nelson argues that the intuition favoring Haydn comes from the "monotony threshold" (infinite mild pleasure becomes boring), not "quality" superiority (Nelson. 2020, p. 105]).

Critics appeal to the resonance constraint (Railton, 1986, p. 9]; Dorsey, 2011, p. 185]): For something to be good for you, it must "resonate" with you-i.e., you must find it compelling or attractive (at least if you're rational and informed). BPH violates this because it says pleasure is good for you even if you don't find it compelling (e.g., a pleasure you don't desire). Nelson refutes the resonance constraint with a coma thought experiment, showing pleasure is still beneficial even without "compelling attitudes" (Nelson, 2020, p. 118]).

In sum, Nelson's work establishes BPH as defensible (surviving descriptive/evaluative objections) and preferable (via simplicity; Nelson, 2020, p. 130]); it is invaluable for scholars engaging with prudential hedonism, as it clarifies BPH's core, provides a template for defending hedonic theories, and highlights simplicity as a decisive theory-selection factor, strongly suggesting its practicability.

4 INSTRUMENTAL OBLIGATION

Since BPH concerns only prudential aims, and prudential aims can be brought about through hypothetical imperatives, I will now investigate a logic of instrumental obligation in order to better formalize BPH. (i.e., "an agent ought to perform an action to achieve a goal"), Yan and He 2025 develop a causal-deontic logic that integrates causal models (to capture action-goal relationships) and priority structures (to rank the desirability of
outcomes). This system rigorously formalizes "what one ought to do to achieve a goal" by integrating causal reasoning (whether an action works) and deontic reasoning (whether it is the best way to work).

Consider, for example, John: John wants to lose weight. But how could he reach that goal? He brainstormed two realistic options. The first was simple: exercise. The second option seemed easier: take weight-loss pills. John is a practical person, so he dug into the details of each choice. He learned that the weight-loss pills would help him shed pounds but come with side effects. On the other hand, exercising had no such downsides, except that it would take more effort-waking up 30 minutes earlier for a walk and skipping a TV show to hit the gym.

The original model formalizes instrumental obligation through three core components: a causal model, intervention-based causal effects, and a priority structure for ideal ordering.

4.1 Causal Model

To John, this model answers: "What variables affect my weight loss, and how do they interact?" The causal model is defined as a tuple $M = \langle S, F, A \rangle$.

4.1.1 Signature $S = (U, V, \Sigma)$

The signature S defines all variables relevant to John's goal, their types, and their possible values (Yan and He 2025, paras. 47,48):

- U: Exogenous Variables-Factors John can't control. For John, these are U_A ("Do I have time to exercise?") and U_B ("Can I buy the pills?")-he can't force himself to have free time or make pills available.
- V: Endogenous Variables-Factors John has the ability (can) to control (or their outcomes). For John, these are:
- A: "Do I exercise?" ($A = 1 =$ yes, $A = 0 =$ no);
- B : "Do I take pills?" ($B = 1 =$ yes, $B = 0 =$ no);
- C : "Do I lose weight?" ($C = 1 =$ yes, $C = 0 =$ no);
- D : "Do I get side effects?" ($D = 1 =$ yes, $D = 0 =$ no).
- Σ (Variable Domain): What each variable can "be." The original model uses $\Sigma = \{0, 1\}$-for John, this means every variable is either "true" (the event happens) or "false" (it doesn't). There's no middle ground, and no measure of how "good" or "bad" the event is.

4.1.2 Structural Functions F

F is a set of rules f that define how variables cause each other. For John, these rules answer: "If I exercise or take pills, what will happen?" The original model writes these
as (Yan and He 2025, paras. 93,97):

$$\{f_A\left(A^{-A}\right) = 1 \Leftrightarrow A\left(U_A\right) = 1\ f_B\left(A^{-B}\right) = 1 \Leftrightarrow A\left(U_B\right) = 1\ f_C\left(A^{-C}\right) = 1 \Leftrightarrow A(A) = 1 \vee A(B) = 1\ f_D\left(A^{-D}\right) = 1 \Leftrightarrow A(B) =$$

Translating this into John's life:

- f_A : "I will exercise ($A = 1$) if and only if I have time ($U_A = 1$)"
- f_B : "I will take pills ($B = 1$) if and only if I can buy them ($U_B = 1$)"
- f_C : "I will lose weight ($C = 1$) if and only if I exercise ($A = 1$) or take pills ($B = 1$)"
- f_D : "I will get side effects ($D = 1$) if and only if I take pills ($B = 1$) "

4.1.3 Actual State *A*

A is the set of current values for all variables (Yan and He 2025, para. 48). For John right now:

- He has no time to exercise ($U_A = 0$);
- He can't buy pills (the pharmacy is out, $U_B = 0$);
- So he doesn't exercise ($A = 0$) or take pills ($B = 0$);
- He hasn't lost weight ($C = 0$);
- He has no side effects ($D = 0$).

In the model, this is written as $A = \left(U_A = 0, U_B = 0, A = 0, B = 0, C = 0, D = 0\right)$.

4.2 Intervention-based Causal Effects

Intervention is the model's way of asking: "If John forces himself to exercise (or take pills), what will happen?" The original model formalizes this as:

$[\vec{V} = \vec{y}]\phi$

This means, "If we force the variables $\vec{V}$ to take values $\vec{y}$, then the outcome ϕ will happen" (Yan and He 2025, para.56). For John, we test two key interventions:

4.2.1 Intervention 1: "What if John forces himself to exercise? ([$A = 1$])"

In the model, this is $[A = 1](C = 1 \wedge D = 0)$. Translating to John:

- Forcing $A = 1$ means John decides to exercise, even if he has to make time (overriding $U_A = 0$).
- From f_C, $A = 1$ guarantees $C = 1$ (he loses weight).
- From f_D, $A = 1$ means $B = 0$ (he doesn't take pills), so $D = 0$ (no side effects).
- Outcome: If John exercises, he will lose weight and have no side effects.

4.2.2 Intervention 2: "What if John forces himself to take pills? ([$B = 1$])"

In the model, this is $[B = 1](C = 1 \wedge D = 1)$. Translating to John Yan and He 2025, para. 127):

- Forcing $B = 1$ means John buys pills elsewhere (overriding $U_B = 0$).
- From f_C, $B = 1$ guarantees $C = 1$ (he loses weight).
- From f_D, $B = 1$ guarantees $D = 1$ (he gets side effects).
- Outcome: If John takes pills, he will lose weight but have nausea.

4.3 Priority Structure for Ideal Ordering

The original model uses a "priority graph" to rank outcomes by how much John should care about them (Yan and He 2025, 67). For John, this answers questions like: "Is losing weight worth having side effects? Is no side effects better than losing weight?"

4.3.1 P-Graph Definition

The P-graph $G = \langle \Phi, < \rangle$ has two parts:

- Φ : The outcomes John cares about-for him, these are "side effects ($D = 1$) ", "no weight loss ($C = 0$) ", "weight loss ($C = 1$) ", and "no side effects ($D = 0$) ".
- <: A "worse than" relation- $\psi < \phi$ means " ψ is less important (worse) than ϕ". (Yan and He 2025, para. 67).

4.3.2 P-Graph for John's Case

For John, the original model assumes a common-sense priority: Yan and He 2025, para. 98)

$D = 1 < C = 0 < C = 1 < D = 0$

Translating this to John's values:

- "Having side effects ($D = 1$) is the worst"-nausea is worse than not losing weight.
- "Not losing weight ($C = 0$) is better than side effects but worse than weight loss"-staying overweight is bad, but better than being sick.
- "Losing weight ($C = 1$) is better than not losing weight but worse than no side effects"-losing weight is good, but not if it makes him sick.
- "No side effects ($D = 0$) is the best"-being healthy and pain-free tops everything.

4.3.3 Ideal World Ordering

Using the P-graph, we rank "worlds" (situations with combinations of variable values) by how well they fit John's priorities. For John, key worlds and their ranks are in Table 1 (Yan and He 2025, paras. 104,106):

Table 1: Ideal Ordering of Worlds in John's Case (Original Model)

World ID	Exogenous Variables (U_A, U_B)	Endogenous Variables (A, B, C, D)	Satisfied P-Graph Propositions	Priority Ranking (Best → Worst)
A_1	$(0, 0)$	(0, 0, 1, 0)	$C = 1, D = 0$	1 (Optimal)
A_2	$(1, 0)$	(1, 0, 1, 0)	$C = 1, D = 0$	1 (Optimal)
A_3	$(0, 0)$	(0, 0, 0, 0)	$C = 0, D = 0$	3 (Current State)
A_4	$(1, 1)$	(1, 1, 1, 1)	$C = 1, D = 1$	4 (Suboptimal)
A_5	$(0, 1)$	(0, 1, 1, 1)	$C = 1, D = 1$	4 (Suboptimal)

To sum up, according to Yan and He, instrumental obligation $O(X = x: \vec{Y} = \vec{y})^{\vec{u}}$ means "to achieve $X = x, \vec{Y} = \vec{y}$ ought to be performed". It satisfies three conditions (Yan and He 2025, paras. 125-126):

$$O(X = x: \vec{Y} = \vec{y})^{\vec{u}} := (\neg X = x)^{\vec{u}}_{\backsim(1)\ GoalUnachieved} \wedge ([\vec{Y} = \vec{y}]X = x)^{\vec{u}}_{\backsim(2)\ ActionCausesGoal} \wedge \Lambda_{\vec{Z} \subseteq V}([\vec{Z} = \vec{z}]X = x \rightarrow [\vec{Z}$$

For John's case:

- $X = x$: John's goal— $C = 1$ (lose weight).
- $\vec{Y} = \vec{y}$: John's action- $A = 1$ (exercise).
- Now check the three conditions:
- $(\neg C = 1)^{\vec{u}}$: "John hasn't lost weight yet"-his current state is $C = 0$, so this is true.
- $[A = 1]C = 1$: "Exercising causes weight loss"-from the intervention, we know this is true.
  $-\Lambda_{\vec{Z} \subseteq V}([\vec{Z} = \vec{z}]C = 1 \rightarrow [\vec{Z} = \vec{z}]\ \vec{U} \leq [A = 1]\ \vec{U})$: "Any other action that causes weight loss is worse than exercising"-the only other action is taking pills ($B = 1$), which leads to side effects (worlds A_4/A_5 which are both worse than A_2). So this is true.

5 MODIFIED INSTRUMENTAL OBLIGATION FOR BASIC PRUDENTIAL HEDONISM

Imagine John again: he weighs 20 kg more than his ideal weight, and his doctor warns him to lose weight. He still has two options: exercise daily (A) or take over-the-counter weight-loss pills (B). But differently, he's unsure if pills cause nausea (side effects, D) and if exercise will actually help him lose weight (C)-he's seen friends exercise without results, and friends use pills without having side effects. He also hates the soreness of
working out but loves the idea of fitting into his old clothes, since he's a hedonist and weighs pleasure over pain. Moreover, John can learn from his experiences.

This is the dilemma of an "ignorant hedonistic agent": John lacks full causal knowledge (ignorance), prioritizes net pleasure (hedonism), and learns from experience. The original causal deontic model fails to model ignorant hedonistic agents due to its deterministic and utility-free assumptions, while also ignoring other hominine factors. To address the original model's limitations, the modified framework extends it in seven key dimensions.

5.1 Utility Extension of Variables and Goals

For John, exercise is painful and weight loss is joyful. To express "how does every variable make John feel?", we extend variables to include hedonic value: a number measuring to which degree does John regard it as pleasure (positive) or pain (negative) for each state. Therefore, the original variable domain $\Sigma = \{0, 1\}$ is extended to a tuple:

$$\Sigma^{H} = \left\{\langle v, h_v\rangle | v \in \{0, 1\}, h_v \in R\right\}$$

h_v (hedonic value) quantifies pleasure (positive) or pain (negative). Generally, when values of variable state is 0 , the hedonic value h_v is correspondingly 0 .

Translating this to John's feelings: we assign h_v based on how he experiences each state (Table 2):

Table 2: Hedonic Values for Variables in John's Case (Modified Model)

Variable State	Description	Hedonic Value h_v	Rationale (Common Sense)
$A = 1$	Exercises	-2	Mild pain from physical effort
$A = 0$	Does not exercise	0	No impact
$B = 1$	Takes pills	-1	Minimal discomfort from pills

$B = 0$	Does not take pills	0	No impact
$C = 1$	Loses weight	+10	Significant pleasure from health
$C = 0$	Does not lose weight	0	No benefit
$D = 1$	Has side effects	-15	Severe pain from discomfort
$D = 0$	No side effects	0	No negative impact

John's goal shifts from abstract $C = 1$ to "maximizing total hedonic value $HV(\phi) = \sum h_v''$ (where ϕ is an outcome proposition).

5.2 Belief-Weighted Causal Model

As said in the case, John doesn't know whether things will definitely occur, he can only make beliefs (with probabilities), denoted as μ. Therefore, the deterministic causal model M is replaced with:

$M^B = \langle S^H, F^\mu, A^\mu, \mu \rangle$

- $S^H = \left(U, V, \Sigma^H\right)$: Extended signature with hedonic values. (From Modification 1)
- F^μ : Set of competing causal rules (capturing ignorance). For John's uncertainty about "whether exercise/pills cause weight loss":
- Rule $f_1: A = 1 \rightarrow C = 1$ (exercise causes weight loss), belief $\mu\left(f_1\right) = 0.6$;
- Rule $f_2: A = 1 \rightarrow C = 0$ (exercise does not cause weight loss), belief $\mu\left(f_2\right) = 0.4$;
- Rule $f_3: B = 1 \rightarrow C = 1$ (pills cause weight loss), belief $\mu\left(f_3\right) = 0.8$;
- Rule $f_4: B = 1 \rightarrow C = 0$ (pills do not cause weight loss), belief $\mu\left(f_4\right) = 0.2$.
- A^μ : Probabilistic assignment to exogenous variables (capturing ignorance), e.g., $\mu\left(U_A = 1\right) = 0.7$ (70
- μ : Subjective probability function quantifying beliefs about rules and exogenous variables.

5.3 Probabilistic Expected Hedonic Value Intervention

The original model says exercise guarantees weight loss-but John knows it might not. We can introduce the concept of expected utility:

$$EV = \sum_{s \in S} \mu(s) \cdot V(s)$$

- μ : belief (epistemological) possibility
- S : all possible cases
- V : value
- $\sum_{s \in S}$: sum of value in all possible cases

Since John's uncertainty lies on the binary relationships, i.e., $f = A \Leftrightarrow B$, and his value is regarding hedonistic value (we can use function $HV(f(\vec{Y} = \vec{y}, U = u) = h_v$ to mark the relationship between intervention and hedonistic value), we can calculate expected hedonic value: the average joy/pain John will feel from an action, while considering his uncertainty. Formally:

$$[\vec{Y} = \vec{y}]HV(\phi) = \sum_{f \in F^{\mu}} \sum_{u \in U} \mu(f) \cdot \mu(U = u) \cdot HV(f(\vec{Y} = \vec{y}, U = u))$$

Translating this to John's question "Will exercise make me happy, on average?", we break down all possible outcomes of exercise, multiply each by how likely it is, and add up the results (Table 3):

Table 3: Calculation of Expected Hedonic Value for Interventions in John's Case

Intervention	Causal Rule	Rule Belief $\mu(f)$	Exogenous Variable U	Exogenous Probability $\mu(U = u)$	Outcome Proposition $f(\vec{Y} = \vec{y}, U$	Total Hedonic Value HV	Weighted Contribution $(\mu(f) \cdot \mu$
$[A = 1]$	f_1	0.6	$U_A = 1$	0.7	$A = 1, C$	$-2 + 10$	$0.6 \times 0.$
	f_1	0.6	$U_A = 0$	0.3	$A = 0, C$	0	$0.6 \times 0.$
	f_2	0.4	$U_A = 1$	0.7	$A = 1, C$	$-2 + 0$	$0.4 \times 0.$
	f_2	0.4	$U_A = 0$	0.3	$A = 0, C$	0	$0.4 \times 0.$
Subtotal	-	-	-	-	-	- | $[A = 1]H$	
$[B = 1]$	f_3	0.8	$U_B = 1$	0.9	$B = 1, C$	$-1 + 10$	$0.8 \times 0.$

	f_3	0.8	$U_B = 0$	0.1	$B = 0, C$	0	0.8×0.1
	f_4	0.2	$U_B = 1$	0.9	$B = 1, C$	$-1 + 0$	0.2×0.9
	f_4	0.2	$U_B = 0$	0.1	$B = 0, C$	0	0.2×0.1
Subtotal	-	-	-	-	-	- | $[B = 1]H$	

5.4 Dynamic Utility-Driven Priority Structure

The original fixed P -graph is replaced with:

$P^{\mu} = \langle \Phi^{H}, <^{\mu} \rangle$

- Φ^{H} : Set of hedonic propositions (e.g., "HV $(A = 1) = 2.8$ ", which means "Exercise gives me +2.8 joy");
- $<^{\mu}$: Ordering based on expected utility (EU): $\phi_1 <^{\mu} \phi_2 \Leftrightarrow EU(\phi_1) < EU(\phi_2)$, which means that ϕ_1 gives less joy than ϕ_2.

For John's case, the priority ordering becomes:

$HV(B = 1) = -7.2 <^{\mu} HV(Inaction) = 0 <^{\mu} HV(A = 1) = 2.8$

5.5 Whole Utility-Oriented Instrumental Obligation

Instrumental obligation is redefined as $O^{H}(maxEU: \vec{Y} = \vec{y})^{\mu}$ ("to maximize EU, $\vec{Y} = \vec{y}$ ought to be performed"), satisfying three utility-focused conditions:

$O^{H}(maxEU: \vec{Y} = \vec{y})^{\mu} := [\vec{Y} = \vec{y}]EU > EU(Inaction)_{\backsim(1)\ Action > Inaction} \wedge [\vec{Y} = \vec{y}]HV > 0_{\backsim(2)\ PositiveNetHedonicValue} \wedge$

Translating to John's dilemma: 1. Exercise's EU (2.8) > Inaction's EU (0): True-exercise makes him happier than doing nothing. 2. Exercise's HV (2.8) > 0 : True exercise gives net joy. 3. Exercise's EU (2.8) > Pills' EU (-7.2): True-exercise is better than pills.

Conclusion: $O^{H}(maxEU: A = 1)^{\mu}$ holds-John ought to exercise, because it makes him happy on average.

5.6 Belief Update Mechanism

John learns from experience. The original model lets John's beliefs stay the same for-ever-but if he exercises and loses weight, he'll be more confident exercise works. We use Bayesian update (Bayes 1763, Prop. 9) to refine his beliefs. Formally:

$$\mu'(f) = \frac{P(Observation|f)\cdot\mu(f)}{P(Observation)}$$

Translating to John's learning: Suppose John exercises ($A = 1$) and loses weight ($C = 1$). He updates his confidence in "exercise works $\left(f_1\right)$ ":

- $P\left(C = 1|f_1\right)$: If exercise works, the chance of weight loss is 80
- $P\left(C = 1|f_2\right)$: If exercise fails, the chance of weight loss is 0
- Original belief: $\mu\left(f_1\right) = 0.6, \mu\left(f_2\right) = 0.4$.

Calculation:

$$\mu'\left(f_1\right) = \frac{0.8\times0.6}{(0.8\times0.6)+(0\times0.4)} = 1.0$$

Now John is 100

5.7 Hedonic Value Update Mechanism

This is when John arrives at a new condition, he suddenly found that, for example, what he thought will bring him pleasure actually brings him pain. It will lead to, in essence, a reassignment of hedonic value:

The core of the mechanism is a weighted average that scales the influence of old and new hedonic values by the number of observations (a common proxy for reliability). The general formula for updating $h^{-}_{X=x}$ is:

$$h^{-t+1}_{X=x} = \frac{n_{old}\cdot h^{-t}_{X=x}+n_{new}\cdot h^{new}_{X=x}}{n_{old}+n_{new}}$$

- $h^{-t}_{X=x}$: old average hedonic value of state $X = x$ (e.g., "jogging occurs"),
- n_{old} : number of old observations (proxy for reliability),
- $h^{new}_{X=x}$: new average hedonic value of $X = x$,
- n_{new} : number of new observations,
- $h^{-t+1}_{X=x}$: updated average hedonic value.

This formula ensures that: If $n_{old} \gg n_{new}$, the old average $h^{-t}_{X=x}$ dominates the update, preventing overreaction to isolated new experiences. If $n_{new} \gg n_{old}$, the new average $h^{new}_{X=x}$ becomes the primary driver, reflecting a robust shift in context.

Translating to John's learning:

- Suppose John loses weight for 20 times, and got an average hedonic value (which is his current hedonic value) of losing weight as 10 . One day it was the 21st time he lost weight, he found the actual hedonic value he experienced is 12 .

Then: Use the weighted average formula to integrate the old 20 experiences and new 1 experience:

$$h_{L=1}^{-t+1} = \frac{n_{old} \cdot h_{L=1}^{-t} + n_{new} \cdot h_{L=1}^{new}}{n_{old} + n_{new}}$$

Substitute the values:

$$h_{L=1}^{-t+1} = \frac{20 \cdot 10 + 1 \cdot 12}{20+1} = \frac{200+12}{21} = \frac{212}{21} \approx 10.095$$

The updated average hedonic value (≈ 10.10) is slightly higher than his old average (which is 10). Reflecting from his 21st weight loss, John found that losing weight was more rewarding than expected.

6 CLARIFICATION OF "OUGHT IMPLIES CAN" IN BASIC PRUDENTIAL HEDONISM

The "Ought Implies Can" principle is often seen as a foundational theorem of moral philosophy ([Kant, 1785, p. 73]); it is intuitive: if you are obligated to do something you cannot do, that is nonsensical-you cannot be required to fly to the moon without a spaceship because you lack the ability.

King ([King, 2014, p. 2]), however, challenges "ought implies can". She presents three different kinds of behaviors: Surface behaviors, like blinking or moving one's arm, which refer to basic, surface-level physical movements; Intentional behaviors, like opening a door or putting on a shirt, which involve clear intentions; Motivated behaviors, like apologizing or genuine sorrow, that is accompanied by a robust mental state and involves an intentional action.

When motivated behaviors become obligations, a mental state is required to fulfill them (though the mental state itself is not obligated; King, 2014, p. 3]); an apology without sorrow is not genuine, and intentions like "kidding" (when apologizing) are forbidden (King, 2014, p. 2]); thus, most behaviors require controlling mental states-but people often lack the ability to control mental states (e.g., a security guard cannot shake distracting thoughts, making "focusing" (a complex action involving mental states) uncontrollable). King argues people "cannot" fulfill such obligations, yet still "ought" to-undermining the principle.

Although she was mainly arguing about moral obligation (King, 2014, p. 1]), her case concerning the nature of behaviors themselves, specifically those in motivated behaviors, might seem to also apply in prudential obligation, e.g., if you are hurt by someone, prudently speaking, you should forgive him, because letting go of hatred will bring you inner peace and greater happiness. But what if your resentment is so deep that you simply cannot forgive from a psychological perspective? Some people would say that even if you can't, you "still should" forgive.

An ethical view (e.g., reasons for action are objective and universal; Benson, 1972, p. 83]) can give rise to resentment, regarding it as the condition happening before the decision of whether to forgive: When others fail to act on reasons that, according to your ethical theory, plainly apply to them - reasons rooted in your plight or interests - you will resent their inaction. For you will judge that your situation itself gives them a reason to modify their behavior, and their refusal to do so violates that universal reason. When you yourself lose control of your mental state and act contrary to such universal reasons (e.g., neglecting your future interests or harming others), resentment toward yourself may emerge as a force to regulate your mental state. This is not merely because you dislike being treated in such a state by others, but because you recognize-from the impersonal standpoint that sees yourself as "merely one person among others"-that your behavior violates reasons that would apply to anyone in that situation. On the other hand, you also expect others to be capable of feeling resentment toward their own actions when they violate universal reasons. This expectation does not arise arbitrarily; it stems from the fact that objective reasons are valid for all persons. If others cannot resent their own violations of reason, they fail to grasp the universality of those reasons - and this expectation thus forms an implicit reason for "one ought to control one's behavior": one ought to regulate oneself to conform to universal reasons, precisely because those reasons bind everyone. The only case where one is truly "out of control" is when one falls into practical dissociation: one can no longer view oneself from the impersonal standpoint, and thus one refuses to act on both prudential reasons (timeless reasons for one's future self) and moral reasons (objective reasons for others' interests). In this state, one severs the conditions for normal interpersonal interaction-for such interaction depends on mutual recognition of universal reasons.

Going back to the case of forgiveness presented by King, initial resentment emerges when others' mistakes are perceived as violating objective reasons-reasons that should motivate anyone to avoid harming others. For example, if someone wrongs us, our re-
sentment embodies the judgment that they failed to acknowledge a reason to respect our interests, a reason that is universally binding . This resentment is rational insofar as it reflects a commitment to the validity of such objective reasons, which altruism "must take as its argument" (i.e., understanding others' self-interest requires first grasping what counts as reasons for oneself). However, because of the impersonal standpoint demanded by altruism, secondary resentment of this initial resentment also emerges. Altruism requires recognizing others as rational beings with their own reasons and limitations, just as we are. When we reflect on our initial resentment, we may realize that we ourselves have similar rational capacity to others'. While reluctant to be resented by others due to limitations of rational capacity, we do not want other people to be resented for the same reason as well. Besides, there's another response to King's "ought doesn't imply can," which is important to remember, regarding that: prudentially, the reasons to act are based on what's in the agent's own interest, and influenced by an individual's beliefs. Across ancient, modern, and contemporary philosophy, resentment has prudential value: it protects the self from exposure to similar actions by deferring people acting in an unjust way in the future, just like the role prisons play in society [Smith, 1853, I.ii.3.5], making resentment useful in the long-term ([Ramacus, 2017, pp. 41-42]);

7 CONCLUSION

This paper's core contribution lies in descriptive decision theory: it describes how ignorant learning agents reason about prudential choices based on Basic Prudential Hedonism (BPH), rather than prescribing normative rules for 'what agents ought to do', which is a task of normative decision theory. The original normative phrasing (ought) in this section is only a logical derivation within the BPH framework, not the core of this study-this paper focuses on describing how agents actually make prudential choices, not prescribing what they should choose. In this paper, I have argued for the following: According to Basic Prudential Hedonism in section 3, we have: II: You should do something if and only if you believe it can enhance your pleasure and mitigate your pain. From section 6, we can know that "ought implies can" can apply to Basic Prudential Hedonism. Then we have: III: If something is your duty to yourself, you must be capable of doing it. ("Ought" implies "can".) These three premises lead to the conclusion that agents will judge an action as prudentially relevant if and only if you believe it can enhance your pleasure and mitigate your pains and your are able to do it. In future research, two research directions can be explored: (1) Developing empirical tools such as fMRI or daily pleasurepain diaries to quantify hedonic value; (2) Extending the model to non-binary choices through multi-variable expected hedonic value calculation, and to sequential decisions via temporal discounting of hedonic value.

REFERENCES

William P. Alston. Pleasure. The Encyclopedia of Philosophy, 6:341-347, 1967.
Aristotle. Aristotle's Nicomachean Ethics. University of Chicago Press, Chicago, IL, 2011. ISBN 978-0-226-02675-6. doi: 10.7208/chicago/9780226026763.001.0001.

Thomas Bayes. An essay towards solving a problem in the doctrine of chances.

Philosophical Transactions of the Royal Society of London, 53:370-418, 1763. doi: 10.1098/rstl.1763.0053.

Paul Benson. Moral obligation and freedom. The Philosophical Review, 81(1):59-81, 1972. doi: 10.2307/2183764.

Ben Bramble. The distinction between mental and physical pleasure. Philosophical Studies, 165(2):201-217, 2013. doi: 10.1007/s11098-012-9903-4.

Franz Brentano. The Foundation and Construction of Ethics. Routledge, London, 2009. ISBN 978-0-415-48751-9.
C. D. Broad. Five Types of Ethical Theory. Routledge, London, 1930. ISBN 978-0-415-28508-5.

Roger Crisp. Hedonism reconsidered. Philosophy and Phenomenological Research, 73(3): 619-645, nov 2006a. doi: 10.1111/j.1933-1592.2006.tb00551.x.

Roger Crisp. Reasonable partiality and the agent's point of view. Ethics, 116(4):725-758, 2006b. doi: 10.1086/505230.

Dale Dorsey. The hedonist's dilemma. Journal of Moral Philosophy, 8(2):173-196, 2011. doi: 10.1163/174552411X571687.

Fred Feldman. Pleasure and the Good Life: Concerning the Nature, Varieties, and Plausibility of Hedonism. Oxford University Press, Oxford, 2006. ISBN 978-0-19-530796-7.

John Martin Fischer. Frankfurt-type examples and semicompatibilism: New work. In Robert Kane, editor, The Oxford Handbook of Free Will, pages 242-265. Oxford University Press, 2 edition, 2012. doi: 10.1093/oxfordhb/9780195399691.003.0012.

Guy Fletcher. A fresh start for the objective-list theory of well-being. Utilitas, 28(2): 148-165, 2016. doi: 10.1017/S0953820815000325.

Immanuel Kant. Grundlegung zur Metaphysik der Sitten. Johann Friedrich Hartknoch, Riga, 1785. Original German edition; References: p. 73.

Immanuel Kant. Groundwork for the Metaphysics of Morals. Yale University Press, 2017. ISBN 978-0-300-12815-4. doi: 10.12987/9780300128154. English translation; References: Section II; pp. 21-23, 39-41.

Amanda King. Actions that we ought, but can't. Ratio, 27(3):316-335, 2014. doi: 10.1111/rati.12047.

Eden Lin. Against welfare subjectivism. Noûs, 50(2):354-378, 2016. doi: 10.1111/nous. 12090.
A. A. Long and D. N. Sedley. The Hellenistic Philosophers. Cambridge University Press, Cambridge, UK; New York, USA, 1987. ISBN 978-0-521-25561-5.

Jon Mandle and David A. Reidy, editors. The Cambridge Rawls Lexicon. Cambridge University Press, 1 edition, 2014. ISBN 978-0-521-19294-1. doi: 10.1017/ CBO9781139026741.

Alfred R. Mele. Free will and luck: Précis. Philosophical Explorations, 10(2):153-155, jun 2007. doi: 10.1080/13869790701305962.

Merriam-Webster. Prudence, 2025. URL https://www.merriamwebster.com/ dictionary/prudence. Accessed: 2025-09-11.

Mark T. Nelson. Basic Prudential Hedonism: A Defense. Routledge, London, 2020. ISBN 978-0-367-35846-9. References: pp. 15, 22, 35, 48, 55, 62, 70, 75, 100, 105, 118, 130.

Peter Railton. Facts and values. Philosophical Topics, 14(2):5-31, 1986. doi: 10.5840/ philtopics198614213.

Lisa Ramacus. Resentment and prudential value. Journal of Value Inquiry, 51(1):41-58, 2017. doi: 10.1007/s10790-016-9563-4.

Stephen Schiffer. Truth and the theory of content. In Herman Parret and Jacques Bouveresse, editors, Meaning and Understanding, pages 204-222. De Gruyter, 1981. doi: 10.1515/9783110839715.204.

Henry Sidgwick. The Methods of Ethics. Hackett Publishing, Indianapolis, IN, 7th edition, 1981. ISBN 978-0-915145-17-2.

Adam Smith. The Theory of Moral Sentiments. John Murray, London, 6th edition, 1853. References: I.ii.3.5.

Dan Sperber. Intuitive and reflective beliefs. Mind & Language, 12(1):67-83, mar 1997. doi: 10.1111/j.1468-0017.1997.tb00062.x.
J. David Velleman. Well-being and time. Pacific Philosophical Quarterly, 72(1):48-77, 1991. doi: 10.1111/j.1468-0114.1991.tb00231.x.

Kadri Vihvelin. Causes, laws, and free will: Why determinism doesn't matter. Ethics, 125(4):1230-1236, 2013. doi: 10.1086/680902. Book review; References: pp. 10-15.

Ming Yan and Jianjun He. A causal-deontic model for instrumental obligation. Journal of Philosophical Logic, 54(2):1-24, 2025. doi: 10.1007/s10992-024-09712-x. References: pp. 12, 15.