

Towards Characterising Conditions of Assistance, Hindrance and Depletion in Human Agency from AI Systems

Tushita Jha^{*1}, Rory Svarc^{*2}, Tom Everitt^{3†}

¹Mimir Center for Long Term Futures Research

²Arb Research

³Google DeepMind

Abstract

Policy and academic discourse around AI often invokes concerns about AI systems’ effects on “human autonomy”. A growing literature also highlights the priority for augmentation, extension, or amplification of human agency. However, as prior authors have pointed out, the concept of ‘autonomy’ is vague and heterogeneous; while on the other hand the character of ‘agency’ that is recommended to be augmented remains in need of deeper philosophical attention. We outline four dimensions for evaluating how artificial assistants may affect autonomous human agency, and argue that these distinct dimensions suggest potentially different practical interventions. More specifically, we distinguish choice availability (the size of an agent’s available action set), volitional integrity (the fit between an agent’s actions and her evaluatively endorsed intentions), referential integrity (the accuracy of beliefs whose accuracy the agent takes to matter), and authenticity (the degree to which evaluative attitudes are the agent’s own). We then analyze the prior literature in light of our four dimensions, and argue that our framework allows for a richer analysis of the potential effects that artificial assistants may have on human autonomy. Then, we apply our framework to related discussions on epistemic agency, recommender systems and on human disempowerment from advanced AI, showing that each conflates analytically distinct agency-impairments. From the framework we derive five maxims for the development and regulation of artificial assistants, thereby centering norms of extension of agency.

1 Introduction

Virginia Woolf told us that there was “no gate, no lock, no bolt that could be set upon the freedom of [her] mind” (1929, pg. 48). Read descriptively, Woolf is close to fully accurate; read as a statement of normative defiance, we agree wholeheartedly. We read Woolf as providing a rallying cry for the importance of human *agency* or human *autonomy*,

where neither gates nor bolts nor AI systems should set limits upon the freedom of human minds. As such, we likewise endorse the more generalized (though less florid) statements offered by various policy documents, stressing that AI systems should respect human agency or autonomy (Université de Montréal 2018; Association for Computing Machinery 2018; OECD 2024; European Commission 2020). At the same time, we are in agreement with the growing literature examining the role of AI systems in augmentation, extension, or amplification of *human agency* (Hernández-Orallo and Vold 2019; Brynjolfsson 2022; London and Heidari 2024; Clark 2025; Jha, Everitt, and Grzankowski 2026). However, both these registers issue promissory notes on the same unarticulated concept: protection presupposes an account of what is being protected, augmentation an account of what is being augmented.

The Problem of Ambiguity. Prunkl notes that policymakers’ use of terms like ‘autonomy’ and ‘agency’ is often unclear. Autonomy (and related concepts) are often deployed in “heterogeneous” ways, leading Prunkl to claim that the status quo risks “groups of researchers and policy makers talking past each other” (2024, pg. 7). Similarly, the growth of the augmentation register highlights the importance of analytic clarity regarding what constitutes the ‘*autonomous exercise of human agency*’ for assessing the proposed mechanisms of augmentation or assistance. If AI development is to serve human agency, it would be helpful to gain greater clarity on what terms like agency or autonomy *actually mean*. Without more precise accounts of agency or autonomy, we risk agreeing to vaporous statements about the importance of AI systems respecting ‘human agency’ without meaningfully steering AI development towards more desirable outcomes.

Philosophical Method. We turn to philosophy of action, or action theory, for the conceptual resources to make these notions more precise. A central strand of this literature has long examined the structure of human agency: the attitudes, capacities, and relations involved in intentional action; and the norms of practical rationality that govern their operation (Davidson 1963; Bratman 1987; Velleman 1992b). In this paper, building on this tradition, in Section 2, we use an en-

^{*}These authors contributed equally.

[†]Author contributed in advisory capacity only. The views, opinions, and regulatory maxims expressed in this paper are those of the authors and do not necessarily reflect the official policy or position of Google DeepMind.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

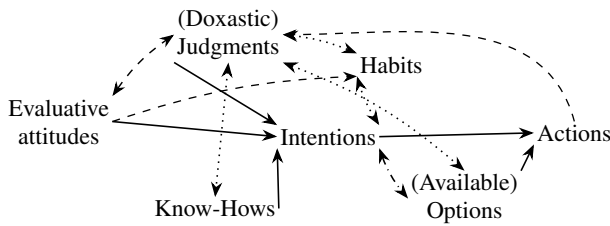


Figure 1: Our schematic model of agency.

riched model of human agency to propose four *dimensions* or *components* – choice availability, volitional integrity, referential integrity, and authenticity – that jointly articulate conditions for the *autonomous exercise of human agency*¹, and provide criteria against which the bearing of any AI system on an agent may be assessed. In analogy with the norms of practical rationality studied in philosophy of action – norms governing the relations among an agent’s intentions and beliefs, and the processes of practical reasoning – we propose, in Section 3, five maxims that function as prescriptive principles for the *well-functioning extension of practical agency*. We also discuss and respond to some possible objections in Section 4, before concluding in Section 5. With this, we thus hope to lay the groundwork for an action-theoretic analysis of how human agency can either be extended, assisted, or augmented; or can be hindered or have its functioning be depleted.

2 Four Components of Agency

Stylised ‘belief-desire’ and ‘belief-desire-intention’ models are common within philosophy of action. The background picture of agency with which we will be operating contains (as illustrated by Figure 1) more component parts. By way of elaboration, we begin by contrasting our picture with the more commonly known *Belief-Desire-Intention* (BDI) model of agency².

In the standard BDI setup, an agent’s intentions arise from her *doxastic* and *evaluative* attitudes. It is commonly accepted that both doxastic states (canonically in the form of beliefs) and evaluative attitudes (canonically in the form of desires) are important components of intention formation,

¹Here, we don’t intend these terms to resolve the heterogeneous uses of ‘agency’ and ‘autonomy’ across the relevant literatures. Rather, we fix a distinction for the purposes of the present analysis, with the broader aim of increasing clarity about the object of inquiry invoked by both terms. We use *agency* to refer to the capacity through which an agent forms and executes action, including the structured relations among the attitudes, dispositions, and practical capacities involved in doing so. We use *autonomy* to characterize the quality of the exercise of that capacity: roughly, the extent to which an exercise of agency proceeds under conditions that permit it to count as the agent’s own and as appropriately self-directed.

²Velleman (1992b) summarizes this ‘standard story’ as: “There is something that the agent wants, and there is an action that he believes conducive to its attainment. His desire for the end, and his belief in the action as a means, justify taking the action, and they jointly cause an intention to take it, which in turn causes the corresponding movements of the agent’s body”

and this paper follows suit. We refer to an agent’s evaluative attitudes simply as ‘evaluative attitudes’, and refer to an agent’s doxastic states as her *judgments*. ‘Judgments’ include straightforward beliefs or degrees of belief such as ‘Paris is the capital of France’, or ‘Arsenal are 40% likely to win the league next year’, alongside broader normative judgments (e.g., ‘I think it is good to keep one’s promises’).³ We incorporate these features in our picture of agency, using the left-right arrow to represent the fact that each may influence the other. Additionally, we allow for the process of intentional action to itself involve potentially revising one’s evaluative attitudes or judgments based on one’s initial set of intentions and options, and similarly denote this relationship with a left-right arrow.

Know-Hows and Habits. Our picture of agency further includes *know-hows* and *habits*, both of which play an important role in the formation of an agent’s intentions. An agent who knows *how* to ride a bike, for instance, is more likely to form the intention to ‘cycle to work’ than an agent without such know-hows. Similarly, an inveterate pessimist may habitually consider the worst-case scenarios in a manner that consequently affects her downstream intentions (imagine an agent with severe anxiety, who no longer considers certain options after years of avoiding situations perceived as risky). Habits⁴ and know-hows are each distinct

³Two notes. First, we distinguish between *evaluative attitudes* (including desires) and judgments, centrally thought of as (factual) beliefs, in addition to potentially normative judgments (e.g., one may have desires one *judges* that one should not act upon). Second, we use slightly more inclusive terminology than ‘beliefs’ and ‘desires’ compared to the standard BDI model. Hence ‘credences’ (as in standard expected utility theory) and ‘risk-attitudes’ (as in Buchak’s (2013) risk-weighted expected utility theory) count as judgments and evaluative states, respectively.

⁴Here we use ‘habits’ to refer to the general class of behaviours that occur without explicit intentions. Hyman (2017) characterizes this general class of behaviors as follows, noting that:

“there are several kinds of exceptions to the generalization that all human action is intentional under some description or stems from an intentional movement, including the following: automatic reactions, such as ducking or drawing back one’s head to avoid a blow, or making an involuntary adjustment to one’s posture to maintain balance; some kinds of habitual action, including verbal tics such as echolalia (the automatic repetition of words and phrases spoken by a person one is conversing with) or interspersing speech with words or phrases like ‘you know’; some kinds of uncontrolled action done in abnormal or pathological states of mind, such as panic or psychosis; unconscious action such as murmuring in one’s sleep; and the spontaneous expression of emotion in facial expressions, vocalizations, and gestures, such as smiling, scowling, pouting, shrugging, and laughing, or crying out with pleasure or pain.”

Similarly, Ouellette and Wood (1998) discuss empirical evidence for a dual process model, suggesting that intentions and habits jointly predict human action. Further discussion of how habits relate to the standard BDI model can be found in (Bermúdez and Felletti 2021; Douskos 2017; Pollard 2008), and Pollard (2006) goes further to argue that habits play a more central role than deliberative action, concluding that “we only think when our habits give

from both judgments and evaluative attitudes – an agent’s *know-hows* are neither judgments about nor evaluations of (the desirability of) bike-riding, and neither are her habits reducible to combinations of judgments and evaluative attitudes. But we think both play an important role in forming an agent’s intentions, and may play an important role in serving or impeding an agent’s degree of autonomy. In particular, when considering the diachronic perspective, recognizing the role of (*dependence on*) *habits* and (*atrophy of*) *know-hows* becomes crucial to provide an account of depletion of human agency.

The Process of Intention Formation. An agent’s intentions may be more or less granular. One may intend to “have fun at the party”, to “introduce oneself to Peter”, or indeed “to inform Peter of the burrito I ate at exactly 12:13pm”. In some cases, fine-grained intentions lead an agent to consider many possible options (e.g., representing many distinct ways to have fun at the party). In other cases, coarser intentions may (more or less) lead directly to action. An agent with more *fine-grained* intentions can be viewed as an agent with a greater number of ‘available options’, if only because such options are more finely individuated in the agent’s representational psychology. Thus, we view the formation of intentions as causally downstream of an agent’s subjective set of *options*, which are themselves influenced by an agent’s beliefs, desires, habits, and know-hows. Acknowledging the highly schematic nature of our picture, the remainder of this section introduces four dimensions of agency motivated by this framework. Each of them can be understood as an integrity condition on some of the elements or links.

2.1 Choice Availability

One feature of human agency involves the experience of *choice* among a set of subjectively available options. We can introduce this idea more formally (if somewhat artificially) by considering an agent A in some world w_1 , faced with a set of available actions $\{x_1, x_2, \dots, x_j\} = X$. As an illustrative example, one may consider an author, who can currently choose among a set of available actions, one subset of which includes {continue writing, make coffee, lift some weights}. When we say that an action x_i is ‘available’ to A , two requirements must be met.

Requirement 1: If x_i is available to A , it is *physically possible* for A to perform x_i .

Requirement 2: The action x_i must show up as *mentally available* to A .⁵

Requirement 1 is intuitive. For instance, if A lacks relevant access to actuators which result in the realisation of x_i ,

out”.

⁵We note that the notion of ‘mental availability’ has some unfortunate circularity. However, we are not attempting to provide a *definition* of the general concept ‘availability’, but rather to highlight two necessary conditions we are imposing on our stipulated concept of ‘availability’ for the purpose of further normative analysis.

then the action is clearly unavailable to them.⁶ The necessity of Requirement 2 can be illustrated by highlighting some examples where this requirement is *not* met⁷. For instance, it may be physically possible for me to move my muscles such that I balance across a slackline even though I lack the know-how to do it. Similarly, an agent may have both the know-how and physical capability to induce political change, but unless they also believe they are capable of doing it, they won’t try. This preamble now allows us to introduce *Choice Availability*.

Choice Availability (CA). If some action x_k is *available* to A in w_1 but not in w_2 — and the remaining actions available to A are otherwise identical in w_1 and w_2 — then A has more *pro tanto* choice-availability (in w_1 than w_2).

Clarifying CA. Increasing an agent’s **CA** should, in general, be expected to increase their autonomous agency. Intuitively, we tend to think that individuals have greater autonomy when they have *more options* — consider, for instance, the way we might think about providing a wheelchair to someone with an ambulatory disability. The wheelchair can be viewed as providing its user with a larger number of available options, thereby increasing their degree of agency. Still, we should note that increasing an agent’s **CA** will not *always* increase their level of autonomy. An agent who is severely misled about the world and the effects of their actions may lack the required understanding to act purposefully, for example. Nor is the addition of an option always a benefit, since giving an agent an option can transform her default position into something she is accountable for choosing (Velleman 1992a). Relatedly, expanded option sets can burden deliberation itself (Iyengar and Lepper 2000; Scheibehenne, Greifeneder, and Todd 2010; Chernev, Böckenholt, and Goodman 2015). **CA** is strictly *pro tanto*: its contribution to agency can be defeated by the cost-profile of the options added.

Choice Availability and AI Systems. First consider concerns about AI-driven ‘human disempowerment’ or human extinction (Bostrom 2014; Dung 2025; Cappelen, Goldstein, and Hawthorne forthcoming). Clearly, total human extinction is the most straightforward example of **CA**-reduction possible. Arguments for ‘human disempowerment’ falling short of extinction are often spelled out less explicitly, though they seem to involve worries about future **CA** violations in at least some cases (Christiano 2019; Kulveit et al. 2025). Second, consider recommender systems. In many cases critiques of recommender systems also involve (de

⁶The sense of ‘actuators’ deployed here is deliberately catholic, intended to encompass ‘actuators’ that may conventionally be thought to belong to another person. For instance, a physically disabled individual lacking voluntary control over the movement of their body’s limbs would count as possessing ‘actuators’ so long as they were accompanied by a physically able-bodied individual who performs their action-requests.

⁷The notion of mental availability connects our account to work on attention as selection for action (Seidman 2009; Wu 2014; Watzl 2017): an option enters deliberation only if attentional processes select it.

facto) claims of CA-reduction, where it is claimed that recommender systems “routinely” provide users “with a limited amount of options”, consequently constraining human choice by “filter[ing] out what is determined to be irrelevant to the user” (Prunkl 2024, pp. 2-4). Reliance upon such recommender systems, thus, can produce a depletion of autonomy through the erosion of CA. On the flipside, the potential normative *promise* of recommender systems often likewise invokes CA-type considerations. For instance, Stray et al. (2024) note that recommender systems “can suggest items without an explicit query, which [...] can lead to positive and novel outcomes for people” (pg. 6).

2.2 Volitional Integrity

The previous subsection noted that an agent who has had their CA reduced (or enhanced) has not necessarily had their agency reduced (or enhanced). By way of illustration, consider the *unwilling addict*. An individual struggling with addiction, who may indeed *welcome* interventions which remove the possibility of engaging in their addictive object; in this case, reducing CA may be, on net, a way of *enhancing* their autonomy.

Motivating VI. There are many reasons we may think that something normatively undesirable occurs in cases like the unwilling addict. We may think the issue arises from conflicts between the agent’s motivational and evaluative systems (Watson 1975, pg. 215), between their first- and second-order desires (Frankfurt 1971), or otherwise think that such cases involve the presence of desires impeding the agent’s ability to self-govern (Bratman 1987; Korsgaard 1996). Of course, there is a strong intuitive case for thinking that CA-restrictions are *prima facie* agency-undermining. However, we may also think that the otherwise agency-undermining effects of CA-restrictions are nullified when *A*’s initial action set interferes with *A*’s evaluative attitudes connecting to her actions ‘in the right way’. A variety of authors within philosophy of action have claimed that (certain forms of) agency require the ‘right kind’ of relationship between evaluative attitudes and physical events,⁸ and this motivates our introduction of *Volitional Integrity*.

Volitional Integrity (VI): If *A*’s actions: (i) depend upon her intentions, *such that* (ii) her intentions are consistent with the subset of her judgments and evaluative attitudes relevant to her current choice of action, *then*: (iii) *A* has more volitional integrity.⁹

⁸We may fail to behave agentically if our actions are not caused by our intentions (Bratman 1987; Bishop 1989; Mele 1992, 2003; Enç 2003), practical reasons (Velleman 1992b; Korsgaard 1996), or appropriate belief-desire pairs (Goldman 1970; Davidson 1963, 1971; Dretske 1988).

⁹Our notion of ‘dependence’ may be read liberally. For readers who believe that some observer-neutral theory of ‘causal dependence’ is important for human agency (perhaps one wants to judge that agents in classic Frankfurt-style cases have their volitional integrity violated under manipulationist theories of causation), ‘dependence’ may be read as ‘causal dependence’. For readers with commitments to the normative value of human agency who reject all such theories, ‘dependence’ may be read as ‘observer-relative

Clarifying VI. We address two sources of potential ambiguity. First, our reference to ‘the subset of *A*’s judgments relevant to her current actions’ is meant to exclude cases where an agent possesses inconsistent combinations of judgments and intentions which are nonetheless irrelevant to her current choice of action. For instance, imagine that *A* is in the process of deciding which dress to buy. If *A* is a realistic human agent, she may have inconsistent judgments or beliefs on a host of topics irrelevant to her dress-purchasing decision. Because *A*’s broader set of judgments are inconsistent, her judgments cannot be consistent with her evaluative attitudes *tout court*. We do not say that *A*’s level of agency is undermined in this setting.

Second, note that condition (i) need not run through explicit, act-by-act intention. For example, consider how a dancer mid-performance does not form a separate intention for each step, but the movements flow habitually. Her VI is nonetheless intact, as the routine itself stands in the right relation to her intentions – she acquired it intentionally, and intentionally initiated it, and remains able to modulate or abandon it. Dependence upon intention can be distal and structural rather than direct and explicit (Shepherd 2021; Wu 2013; Mele 1992; Bratman 1987).

Third, our concept of VI also allows us to better evaluate the degree to which an addict’s autonomy or agency is impeded. Her autonomy will be more impeded in cases where she explicitly *judges* (or believes) that engaging in addictive behaviors is inconsistent with her *intended way of life* but does so anyway due to weakness of will (e.g., “it is a bad idea for me to take this”) – as this would illustrate an inconsistency between her ideals for self-legislation and what she actually does. Addictions can be thus understood as deeply ingrained *habits* leading to unwanted actions. Whether such unwanted actions undermine volitional integrity depends on whether or not these habits were voluntarily adopted in the first place, and whether—once akrasia is noticed by the agent—it can be modified through voluntary intentions. When adequately empowered, agents should be able to modify their automatic behaviour habits in greater concordance with their values.

Volitional Integrity and AI Systems. Because recommender systems “track only individuals’ first-order preferences or desires by surveilling their actions”, Laitinen and Sahlgren (2021) claim that such systems may be “particularly detrimental to autonomy” in cases of “addictions or akratic will”. This component of Laitinen and Sahlgren’s analysis can clearly be viewed as a threat to human VI. Under the heading of a closely related concept, Brown claims that “recommender systems foster volitional misalignment by design as they rely on behavioral signals that poorly correlate with their users’ second-order desires” (Brown 2026, pg. 23). Using our terminology, Brown is concerned with VI-degradation. Outside of recommender systems, Ashkinaze and co-authors explicitly test (what they term) LLMs’

evidential dependence. For instance, we may say that *A*’s actions depend upon her evaluative attitudes when knowledge of her evaluative attitudes increases her probability that she will perform the actions she in fact performs.

“deep value generalization” on human preference-data involving correlations between first- and second-order desires, and find that “current LLMs [as of 2025] predominantly favor shallow preferences” (Ashkinaze et al. 2025, pg. 10). To the extent this finding holds more generally and robustly, this would be a sign that current LLMs may have negative effects on humans’ VI.

2.3 Referential Integrity

We have so far seen two concepts relevant to a person’s agency. Namely, we could increase (decrease) their set of available actions, or increase (decrease) the degree to which they reflectively endorse their actions. An additional way in which we could enhance a subject’s agency involves *informing* them of certain action-relevant facts. A classic case from Williams (1981) serves as our initial illustration.

Williams’ Case and Analysis. In the standard setup, an agent (*A*) is motivated to drink the translucent liquid (petrol) in the glass in front of her, on account of her belief that the object in the glass is gin. Because *A* believes that the liquid is gin, she would act in a *subjectively rational* manner by drinking the liquid. If an outside observer knew these facts and withheld accurate information from *A*, this would plausibly undermine *A*’s autonomy. On account of *A*’s ignorance, she lacks the effective capacity to act in a way that fulfills the content of her intentions; providing *A* with knowledge thereby enhances her ability to serve her present desires. Thus we define:

Referential Integrity (RI). Suppose that *A* possesses some doxastic attitudes. *A* has more *Referential Integrity* when, from *A*’s perspective, her doxastic attitudes *in fact* bear the appropriate relationship to the real world.

Clarifying RI. When we say that *A*’s judgments (from her perspective) in fact bear the appropriate relationship to the real world, we are making two important stipulations.

1. First, we consider *A*’s perspective on judgments whose accuracy matters most to *A*. This is an **internal** matter relating to *A*’s (dispositional) psychology.
2. Second, we ask whether those judgments are *in fact* more accurate. This is an **external** matter relating to the objective accuracy of *A*’s judgments.

We take the normative importance of **RI** to be highly intuitive. In general, we think that we impair or enhance an individual’s **RI** as an increasing function of the importance they would attach to possessing accurate judgments about the matter in question. If we were to ask Williams’ fictional victim whether accurate beliefs regarding whether the liquid was gin or poison mattered to her, she would presumably say ‘yes’, and presumably this would matter a great deal. In more realistic cases involving (for example) addicts, one might assess the degree to which their addictive habits arise from a lack of *self-knowledge* about the influence of the addiction on their life, or how easy it would be to remove themselves from their addiction (“I can quit when I want”).

Furthermore, the capacity to responsibly form intentions and make choices relies upon holding valid judgments about

the consequences (of one’s actions) (Fischer and Ravizza 1998; Bratman 2007; Sher 2009). We may see hindrance to such knowledge as producing, over time, a depletion of agency in such cases¹⁰.

Referential Integrity and AI Systems. A small literature has brought forth worries about technologically abetted depletion of epistemic agency (Lynch 2016; Nguyen 2020; Gunn and Lynch 2021), though a recent paper by Coeckelbergh (2026) bears particular relevance for our discussion. Coeckelbergh’s paper focuses centrally on a threat that future AI systems may pose to humans’ *epistemic agency*, which “concerns the control that agents may exercise over their ‘beliefs’ ” (pg. 1, quoting Schlosser (2019)). More localized variants of this worry appear in recent empirical work (see Kosmyna et al. (2025)). One particularly notable illustration of potential **RI** violations can be found in an example courtesy of OpenAI, who – after publicly declaring an ostensible “improvement” to the ‘personality’ of the model GPT-4o – quickly rolled back their updates, after many users anecdotally reported instances where the model delivered unhelpful (and in some cases, potentially dangerous) sycophantic advice, in one case encouraging a user with transparently delusional beliefs to abstain from their medication.¹¹

2.4 Authenticity and Meaning

We now introduce our final normative concept by first imagining the extended lifetime of an agent *A*. Specifically, we’ll suppose that *A* has led a seemingly salubrious life; her space of available actions has increased, and so too her judgment over (or knowledge of) claims whose accuracy she takes to be important. Her evaluative attitudes have also become more consistent, and her actions more dependent on these evaluative attitudes. Yet (as the astute reader may have inferred from the presence of this subsection), we believe this nonetheless leaves out one normatively important component of *A*’s agency. We thus define:

Authenticity and Meaning (AM). *A*’s actions arise from evaluative attitudes and dispositions which are ‘authentically their own’.

Clarifying AM: The Problem of Coercion. Consider cases of blackmail or coercion. Agents who have been coerced or blackmailed may be acting in full knowledge, and choosing actions which are meaningfully shaped by their evaluative attitudes. Moreover, agents subject to coercion or blackmail may very well come to have an *expanded* set of available actions on account of being threatened or blackmailed. We may suppose that the agent, prior to being blackmailed, had some set of available options. After being blackmailed, they now have the enticing option of ‘giving my life savings to the blackmailer’, which (let’s suppose) was *not*

¹⁰Nozick’s (1974) experience machine can be read as the limit case of **RI** violation: the machine-dweller’s judgments are systematically illusory. In this sense, the repugnance of experience machines can be understood in terms of **RI**.

¹¹See (Williawa 2025) for the original source of the medication response, and (Edwards 2025) for wider discussion before OpenAI tried to reduce this tendency more explicitly.

previously in their available option set. The blackmailee may view all of her previous actions as mentally available — except that, now, her previously available actions come with further costs. We believe there to be a clear intuitive case for thinking such cases involve limits imposed on one’s agency.

Clarifying AM: Inauthenticity. Even absent blackmail or threats, one may further worry that *A* may fail to act *agentially* insofar as she fails to act *authentically*. Christman’s (2009, pg. 155) account of authenticity, for example, requires that *A* feels “non-alienated” from her attitudes and commitments, such that she accepts such commitments as “an ongoing part of [her] ‘autobiographical narrative’”, as organised by her “diachronic practical identity”. Alongside Christman, various authors have claimed that developing properly authentic values requires narrative construction in some form (MacIntyre 1982; Taylor 1989; Schechtman 1996). While we do not mean to suggest the views above are without contest (see, in particular, (Strawson 2004) for a forceful criticism of narrative identities), our definition hopes to avoid strong commitments to any *particular* account of meaning or authenticity, while acknowledging that there are intuitively reasonable conceptions of ‘authentic’ values which our framework would otherwise omit.

AM and AI Systems. Return to Coeckelbergh’s paper on AI and epistemic agency. Coeckelbergh assumes that “Claire wishes to get rid of her belief” (pg. 7), and indeed claims it “plausible” that Claire “wants to be *the kind of person* that does not hold [her current] beliefs” (pg. 8). The scenario described also appears to violate Claire’s **VI** — her actions, such as they are, fail to be consistent with her evaluative attitudes. Additionally, Coeckelbergh’s scenario references to external actors with “the power to promulgate their beliefs”, who use their power to maintain the situation where Claire suffers from reduced “epistemic agency” (pg. 9). This appears to be a violation of Claire’s **AM**, or the degree to which her evaluative attitudes that are “properly her own” are actionable and effective.

Manipulation, or influence that bypasses rather than engages the agent’s capacities for reflective evaluation (Noggle 1996; Susser, Roessler, and Nissenbaum 2019), offers a clear example as it can degrade **AM** even while preserving or expanding **CA**.

More generally, Brown’s discussion of algorithmic filtering notes ways that such systems can *enhance* an agent’s authenticity through “coming to see that [they] behave similarly to people belonging to category *x*”, which Brown dubs as “a significant way in which a person can come to understand who they are” (2026, pg. 25), while also noting that certain implementations involving too much user control over recommender systems may come at the cost of “diminish[ing] the likelihood of encounters that prompt users to question or revise their current desires” (2026, pg. 30), thereby potentially weakening **AM**.¹² Similarly, an AI assis-

tant optimized on who the agent has historically been may impede who she ‘*aspires*’ to become (Callard 2016, 2018).

2.5 Distinct Risks: Analysis and Intervention

We have outlined four normatively important components of human agency, and discussed these concepts in relation to prior literature on the effects of (present or future) AI systems. We now close with a brief discussion pertaining to the usefulness of our proposed concepts; we claim that our framework allows us to disambiguate—and thereby better evaluate and intervene upon—distinct risks raised by previous authors.

Epistemic Agency. One of the papers discussed earlier was Coeckelbergh (2026). While Coeckelbergh’s paper is squarely focused on ‘epistemic agency’, our preceding analysis highlights ways in which his central example—a case involving the character ‘Claire’—plausibly has many distinct components of her autonomy undermined. As already discussed, Coeckelbergh’s case involves components of **RI**, **VI**, and **AM**. For this reason, we consider Coeckelbergh’s normative analysis to be, while helpful, nonetheless importantly incomplete, and we believe that our concepts allow for a more complete normative analysis. In practice, the different facets of autonomy Coeckelbergh discusses under the heading of ‘epistemic agency’ obscure potentially different technical challenges with training AIs to enhance ‘epistemic agency’ broadly construed.

With respect to **RI**, one might think that the presence of attribution bias observed in some LLMs (Suzgun et al. 2025)—where models are more accurate at assessing false beliefs stated in the third-person than false beliefs stated in the first-person—poses a threat to **RI**. Hence, one intervention to increase **RI** might involve reducing attribution bias in LLMs. However, well-known claims of ‘LLM sycophancy’ plausibly arise from the fact that users (at least locally) prefer model responses which validate initial user claims; indeed, there is *also* some empirical evidence that users prefer LLM responses which state confident falsehoods over explicitly expressing uncertainty (Bar-Or Nirman, Weizman, and Azaria 2025). Ensuring that users feel ‘non-alienated’ from any beliefs arrived at through interaction with LLMs may thus require fine-tuning LLMs to provide ‘defeaters’ explaining the falsity of any original beliefs to relevant users. This intervention would, in effect, involve training LLMs to be sensitive to third-person/first-person expressions of beliefs, meaning that naive interventions aimed at increasing **RI** could have corresponding deleterious effects on **AM**.

Recommender Systems. As discussed earlier, Prunkl’s analysis of recommender systems involved (what we are terming) critiques of recommender systems on grounds of **CA** (2024, pp. 2-4). In addition, her paper later draws a connection between issues with recommender systems and the objectionability of “adaptive preferences”, claiming that recommender systems may constrain human autonomy by leading to “alienation and/or ... changes in preferences or self-conception” (2024, pg. 10). The first of Prunkl’s objections appears to reference the **CA** component of human autonomy (‘limiting choices’), whereas the second appears to touch on

¹²Note that we here refer to Brown’s analysis of ‘authenticity’ viewed in light of our definition of **AM**. The definition of authenticity Brown uses in the cited paper is much more widely-scoped, encompassing concerns we label as **VI**.

AM ('alienation'). While Prunkl helpfully distinguishes between 'internal' and 'external' dimensions of autonomy, she treats the former as synonymous with 'authenticity' (2024, pg. 8), leaving the 'external' dimension to capture all remaining factors. However, we have seen the external dimension admits further decomposition still.

In practice, we think our distinctions within Prunkl's conception of 'external autonomy' may be important, as our preceding discussion has shown authors raising concerns about recommendation systems on grounds of both **CA** and **VI**. Even if we restrict ourselves to recommender systems' effects on 'external autonomy', we may well think that concerns about **CA** and **VI** suggest different interventions. In the former case, we may wish to make recommender systems more 'exploratory', so that they do not lock-in on a shallow representation of users' preferences too readily. In the latter case, we may indeed want recommender systems to be in certain contexts *less* exploratory, so that recommender systems do not (as they may well do (Murray et al. 2022)) provide inappropriate recommendations to recovering addicts.

Human Disempowerment. Consider concerns about AI-abetted 'human disempowerment'. Here, '*disempowerment*' standardly denotes scenarios where "humanity is unable to determine its own future", and Dung notes that humans are disempowered when some other entity "[has] control over them, determining what they can and cannot do" (Dung 2025, pg. 1196). Additionally, "catastrophic risk" is used to describe scenarios "involving globally devastating harm or threats to human existence, autonomy or potential" (Bales, D'Alessandro, and Kirk-Giannini 2024, pg. 2). Given that many discussions within this literature wish to make claims about risks to human agency beyond mere human extinction (a straightforward **CA**-violation if there ever were one), future research within this tradition would do well to focus on the more specific threats to human agency envisioned.

If future AI systems 'disempower humanity' or cause 'catastrophic risks', we may wish to know whether humanity's inability to "determine [its] own future" would consist in the existence of fewer options (**CA**), whether future human actions will be less shaped by their evaluative attitudes (**VI**), whether humans are led to false judgments (**RI**), or whether such scenarios primarily involve the absence of meaningful choices (**AM**). Plausibly, different views on *how* humanity gets disempowered by AI will have downstream consequences for these distinct components of autonomy, and hence (plausibly) different views will thereby suggest different interventions for how to protect human autonomy. While work within this tradition is (to some degree understandably) light on detailed threat-models, specifying concrete accounts of how non-extinction variants of human disempowerment may occur would allow us to better assess the specific kind(s) of disempowerment such authors find most plausible and concerning.

In our view, the four dimensions offer a step toward a synthetic program: one that distinguishes the kinds of risk to human agency, and marks the constitutive conditions that well-functioning assistants (or augmentations) should strengthen.

3 Maxims for Assisting and Guiding Intentional Capacity (MAGIC)

This section moves from diagnostics to prescription; we introduce five maxims and claim that they ought to govern the development and regulation of artificial assistants. By analogy with Grice's (1975) discussion of maxims as articulations of what it is for conversation to function well in its communicative context, the prescriptive norms proposed below articulate what it is for assistance and extension to function well. We begin by outlining three 'unconditional maxims' (Section 3.1), and then introduce two further maxims which reference 'quasi-mental states' (Section 3.2). We close by providing a brief defense of our maxims (Section 3.3).

3.1 Unconditional Maxims

The first of our unconditional maxims is the *Maxim of Empowerment*, which can be seen as a direct analogue of **CA**.

Maxim of Empowerment

Maxim of Empowerment (ME). AI systems should weakly expand, not constrain, the user's range of available choices and opportunities.

Intuitively, increasing (decreasing) the set of actions available to an agent was *prima facie* autonomy-enhancing (autonomy-undermining). Insofar as we desire to promote **CA** in the case of humans, we want future AI developments to increase (rather than decrease) the space of options available to human agents. Although one of course may imagine artificial assistants that constrain the available choices of individuals without thereby hindering their autonomy (consider the case of suffering from addictive desires or akratic impulses), we nonetheless think that **ME** constitutes a *pro tanto* burden on AI developers. To the extent that AI developers *fail* to heed **ME**, then AI developers take on a legitimate justificatory burden.

Maxim of Visibility

Our second maxim, meanwhile, can be viewed as a prescriptive analogue of **RI** – the claim that a putative agent may have their agency enhanced (or undermined) by making their beliefs more (or less) accurate.

Maxim of Visibility (MV). Artificial assistance weakly improves humans' awareness of the (subjectively normatively relevant) effects of their actions.

Consider the case where **MV** fails. If artificial assistance degraded humans' awareness of the subjectively normatively relevant effects of their actions, then artificial assistants would be (in our terminology) violating humans' **RI**, or (in other words) making humans less knowledgeable about the subjectively relevant effects of their actions. We do not deny that there may be cases where violating **MV** is desirable (though we take this to be less straightforward than the analogous case for **ME**). Analogous to our discussion of **ME**, however, we take there to be a strong *pro tanto* argument for heeding **MV**.

Maxim of Regulation

Our third and final unconditional maxim will benefit from some preemptive unpacking. When introducing **VI**, we made reference to the consistency of an agent's actions with their evaluative attitudes. And, as discussed, one common avenue by which **VI** is impaired arises in the context of addiction or akrasia. This is to say that agents' volitional integrity is often impaired by their *inability to self-govern*. The capacity to exercise resolutions (Holton 2009) or commitments (Elster 2000) enables such self-governance.

Maxim of Regulation (MR). Artificial assistance extends human self-governance through the externalisation of commitment devices.

Although one may view adherence to **MR** as supererogatory (perhaps AI developers should simply be able to develop 'assistants' which behave as requested), we nonetheless believe that **MR** embodies a desirable normative ideal. Indeed, one may reasonably regret the relatively thin normative ideals common to literature on "the alignment problem" (Russell 2019; Gabriel 2020), which the latter defines as the problem of ensuring AI systems "pursue the values and priorities of the people they are intended to serve" (p. 414).

Rather than enjoining more generally that artificial assistance ought to pursue the "values and priorities" of those the system is intended to serve, **MR** provides (admittedly still incomplete) guidance on which "values and priorities" AI systems ought to be pursuing in cases where an agent may face conflict between different desires or motivations. While we do not claim that **MR** is the final word on whose values artificial assistants ought to pursue (we have not touched on issues of fairness or justice, for example), and do not claim that even our formulation of **MR** always delivers verdicts about what AI systems should do, we suggest that this provides a helpful first attempt at prescribing what it means for an artificial assistant to pursue the values and priorities of the user or group it is intended to serve.

3.2 Conditional Maxims

We now introduce two further conditional maxims, departing slightly from the one-to-one mapping from dimensions to maxims. The justification for this slightly unordered departure stems from concerns that future artificial assistants may possess 'belief'- and 'intention'-like states. Consequently, the two forthcoming conditional maxims will make free use of 'quasi-belief' and 'quasi-intentional' states as applied to artificial assistants. (See Section 4.1 for a response to one possible objection on this point).

Maxim of Service

The first of our conditional maxims is the Maxim of Service, which imposes a relatively weak requirement on forms of artificial assistance.

Maxim of Service (MS). If we develop artificial assistance with quasi-intentional states, then such states should have a means-end relationship to some intentional states of the human users with whom it interacts.¹³

¹³See clarifications of this formulation offered in Section 4.1.

Suppose that we develop artificial assistants with 'quasi-intentional states', which fail to bear a means-end relationship to any intentional states of humans. At best, this is a world where future AI systems are indifferent to human intentions. At worst, this leaves us in a world where AI systems' quasi-intentional states bear a merely instrumental—and potentially adversarial—relationship to human intentions. Violating **MS** indeed appears to be a straightforward failure to heed even the thin normative ideal behind "the alignment problem" referenced earlier (Russell, 2019; Gabriel, 2020). Even when some of a user's intentions rest on mistaken judgments, appropriately provided assistance serves some *deeper* intentional state—the one the mistaken request was itself in service of. We thus consider this maxim to be unobjectionable. Adhering to **MS** does not tackle many possible AI threats, though we believe that it provides a plausible *necessary* condition for the development of genuinely agency-enhancing artificial assistants.

Maxim of Consistency

Finally, we introduce the *Maxim of Consistency*. In the context of our suggested framework, **MC** may be seen as a prescription intended to ensure that artificial assistants promote humans' volitional and referential integrity.

Maxim of Consistency (MC). If the quasi-belief states of artificial assistance judge a human's beliefs or intentions to be inconsistent (or mistaken), then the human is proactively informed of such inconsistencies.

We invite the reader to consider a situation where the model has quasi-belief states which indicate that the human with whom they're interacting possesses inconsistent intentions, and does not proactively inform the human of this belief. At the very least, this involves an AI with quasi-belief states which – either due to system constraints from its developers, intentional behaviour, or architectural quirks – fails to inform the human of one way to improve the coherence of their intentions. If we are to help ourselves to more intentional language, it may even appear as though the artificial assistant is *withholding* information from the human. Hence, the assistant or its developers appear complicit in a violation of the human's **RI**, and very plausibly their **VI**.

3.3 Why Endorse the MAGIC Maxims?

We close this section with a brief justification of the maxims introduced hitherto.

Justifying the Unconditional Maxims. Start with the unconditional maxims. **ME** is a direct prescriptive analogue of **CA**, and there is at least a *prima facie* case for thinking that greater **CA** begets greater levels of autonomy. **ME** simply formalizes this important component of human agency, and our statement of this maxim simply makes it salient in normative discussions relevant to the future of AI development. Similar justifications apply to **MV** in the context of **RI**. With respect to **MR**, this too has its own foundation in the concept of **VI**. We include in our statement of **MR** reference to *extending* human self-governance because we believe it fruitful to consider ways in which AI assistance of-

fers opportunities for augmenting human agency rather than simply defending against potential risks to human agency. This positive extension has roots in prior discussions of normative ‘alignment’ ideals for AI systems, and we believe it has value in virtue of explicating a minimal ideal which can be used to assess the agency-enhancing properties of both present and future AI systems.

Justifying the Conditional Maxims. For a more specific defence of quasi-mentalistic language, see Section 4.1. Regarding the more specific content of our conditional maxims, we take them to be fairly minimal. In the case of **MS**, we encode (what we take to be) a substantive though fairly inclusive conception of ‘AI liberalism’, directly targeted at ensuring that artificial assistants afford humans some degree of autonomy over which goals they would like to pursue. The conception of **MS** we put forward allows for the possibility of systems which refuse requests to (for example) engage in dangerous or dual-use forms of biological research. Our formulation simply wishes to rule out cases where AI assistants are developed with the sole purpose of furthering some external aim. More precisely, then, **MS** should be read as a claim about artificial assistants’ quasi-intentions bearing a means-end relationship to some *de dicto* human intentions.¹⁴

Regarding **MC**, our earlier discussion of Williams’ gin-drinking example in the context of **RI** highlights the importance of being *proactively informed* about errors when considering how to promote and enhance human agency. By proactively informing Williams’ subject that they are about to drink poison rather than gin, we thereby increase their degree of agency.

4 Objections and Replies

This paper has introduced four components of human agency, and suggested five maxims motivated by our framework. We close by responding to potential objections.

4.1 Objections to Mentalistic Language

One may raise philosophical quibbles with our conditional maxims on account of their using ‘quasi-mentalistic’ language. We offer two brief responses.

Non-Ideal Theory. References to the ‘goal-like’ and ‘belief-like’ states of contemporary AI systems are discussed or assumed both within the philosophical literature (Bales, D’Alessandro, and Kirk-Giannini 2024; Bales 2025; Herrmann and Levinstein 2025), and in various papers published by leading AI labs (Greenblatt et al. 2024). As many

¹⁴When an artificial assistant’s quasi-intentions bear a *de dicto* means-end relationship to some set of human intentions *I*, the assistant quasi-intends to help the human achieve *I* — whatever *I* happens to be. By contrast, an artificial assistant’s quasi-intentions may bear a *de re* means-end relationship to *I*, simply because the human’s intentions happen to overlap with the quasi-intentions of the artificial assistant; if the human did not intend to (say) help a misaligned AI system autonomously self-replicate, the assistant’s quasi-intentions would not bear a means-end relationship to the human’s intentions. See (Nelson 2024) for more on the *de re/de dicto* distinction.

AI developers and regulators are using mentalistic language *anyway*, we believe that guidance should be offered regarding the appropriate relationships such assistants should bear to human users. If one maintains philosophical objections to such language, the antecedent of our conditionals will be false and the maxims will be heeded trivially. And if artificial assistants do (in some sense) possess the relevant mentalistic states, we believe that our maxims provide helpful guidance on how these states should relate to the mentalistic states of human users.

Philosophical Innocence. Both Schwitzgebel (2023) and Chalmers (2026) offer novel terminology with which to refer to the computational states of contemporary LLMs. Chalmers coins the concepts of ‘quasi-belief’ and ‘quasi-desire’, which may be ascribed to a system *S* if *S* “is behaviorally interpretable as believing [or desiring] that *p* according to an appropriate interpretation scheme” (2026, pg. 4). We thus use Chalmers’ quasi-mentalistic ascriptions, noting that our use of such language only commits us to believing that these terms may be useful “in predicting a system’s behavior”, such as in cases where attributing a specific quasi-desire for outcome *O* given a set of quasi-beliefs allows us to better predict the outputs of that system as oriented towards *O* (2026, pg. 5).

4.2 Objections to Our Concepts

Another objection might be that we have not attempted to provide necessary and sufficient conditions for someone to count as autonomous. In at least some cases, we acknowledge that agents may have their autonomy enhanced by having their available choice set restricted, or indeed by making an individual’s beliefs less accurate. Hence, one may worry that we have introduced four concepts (we assume even the skeptical reader will grant this much), but think it unclear *why* we have introduced these concepts, or what value (either epistemic or practical) is gained by their introduction.

Conceptual and Practical Value. Although we do not provide necessary and sufficient conditions for autonomy, Section 2 put forward concrete claims. For instance, we claim that each concept introduced in Section 2 constitutes a *prima facie* threat to (or enhancement of) autonomy, and we believe that conceptual clarity is gained by introducing our concepts even if they admit exceptions. As previously noted, ‘autonomy’ is used in myriad ways within the AI policy literature. Just as we believe Prunkl (2024) offers a useful clarificatory intervention by distinguishing ‘internal’ and ‘external’ components of autonomy, we likewise believe that our concepts provide useful tools for both policymakers and social critics with respect to analysing potential AI threats (or enhancements) to human autonomy. Even if (say) restrictions to **CA** are not always autonomy-undermining, one may reasonably presume that **CA**-restrictions are *prima facie* autonomy-undermining. So long as our concepts allow individuals to analyse AI-enabled threats to human autonomy with greater specificity, we consider our framework to have provided value. In contexts where potential concerns about **CA**-restrictions are raised, we may hence expect companies producing artificial assistants with the effect

of restricting human choice to face a justificatory burden. In principle, this could be used (for example) to claim any observed instances of **CA**-restriction must be accompanied by concomitant defences of how human autonomy is being enhanced via other routes (e.g., via improvements to humans' **VI**).

4.3 Objections to Our Maxims

Even if our concepts are coherent in principle, one may believe our maxims to be empty. Channelling Munn (2023), one may believe that three pitfalls befall nearly all work in AI ethics. First, Munn claims principles in AI ethics can be “**meaningless**”. As the Montreal principles helpfully illustrate, principles which refer to the “well-being of all sentient creatures” come with ‘immediate problems’ when we start to define what such terms mean, as there exist trade-offs between the well-being of different sentient creatures (2023, pg. 870). Second, such principles can be **isolated**, insofar as they fail to “encompass the ethical, moral, and social impacts of designing a product in a particular way”, in a way that is “situated downstream from company culture ... [while failing] to address the more fundamental inequalities and underlying social issues that shape technological development” (2023, pg. 871). Finally, such principles can be **toothless**, “not just” when they lack penalties for violation, but when there exists a “lack of friction between ethical principles and existing business principles” (2023, pg. 872).

Response to Meaninglessness. Taking each point in order, we start by noting that our maxims are more ‘meaningful’ than the illustrative statement Munn offers. We think there are clearer concrete application conditions than the example offered. For instance, we note (for **MV**) that many empirical articles have already tried to investigate the cognitive effects of interacting with contemporary AI assistants (Hartley, Hayak, and Ko 2024); one could imagine similar such studies aimed at investigating whether (or the degree to which) humans understand the consequences of their actions after interacting with AI assistants. If there were reason to expect **MV** violations to be particularly severe or widespread, (even informal) assessments of this kind could become part of pre-deployment AI evaluations. Or, with a less regulatory heavy hand, AI companies themselves could run such tests in order to showcase the beneficial effects of interacting with their AI systems.

Regarding **ME**, policymakers could fund surveys asking those who interact with artificial assistants whether they feel as though their range of available options is expanded, or whether (for **MR**) AI tools aid their ability to self-govern. For instance, surveys of users interacting with AIs for the purpose of managing multiple competing life-tasks (e.g., budgeting time for both job applications and childcare), and retrospective surveys of users with addictions could provide evidence on the degree to which such users are aided in their ability to self-govern. Because the maxims we have provided allow for future empirical work to test potential maxim violation, and because the maxims we provide allow for clear circumstances in which the maxims would count as violated, they are not ‘meaningless’ in Munn’s sense.

Response to Isolation and Toothlessness. On the question of ‘isolation’, we simply do not take the ‘isolation’ of a maxim to be dispositive of its value. More importantly, however, as our maxims have *not* (as of the time of submission) yet been published, the critique Munn offers here is, at best, premature. Before evaluating the ability of our maxims to address such issues, they first must be put into the wild. In any case, we put forward requirements on AI assistants to *aid human self-governance*; this is at least plausibly in conflict with more predatory business practices which aim to exploit users’ unendorsed first-order desires. On ‘toothlessness’, we can say a little more. Our maxims are not (at least we hope) entirely in conflict with existing business practices. In principle, however, such maxims could be in conflict with orthodox business practice, which we take to be sufficient for establishing that the maxims are not ‘toothless’. Similar to the Gricean conception (Grice 1975), the maxims articulate the constitutive norms of a practice. And we simply claim that our proposed maxims are a useful first-step towards future work that is sufficiently precise to allow—given further development—for *potential* regulatory guidance and penalties in the future.

5 Conclusion

This paper began by observing that two registers of contemporary AI discourse — one urging the protection of human agency, the other its augmentation — produce a common demand for an actionable account of *autonomous exercise of human agency*. To meet this demand, our paper offered a quadripartite decomposition of human agency in terms of four concepts (Section 2), and thereafter introduced five maxims for AI development based on these four concepts (Section 3). We claim that: (i) the concepts we offer provide a helpful lens with which to analyze and decompose the potential effects of AI systems on human agency, and (ii) that the maxims we suggest provide helpful (though incomplete and preliminary) guidance on how to ensure that AI systems enhance rather than degrade human agency. Our analysis so far has been confined to individual agents; whether collectives possess analogues of the four dimensions (Bratman 2014; List and Pettit 2011; Velleman 1997; Searle 1990; Tuomela and Miller 1988; Gilbert 1987), we leave open. Section 4 provided a defense of our framework in the course of responding to objections concerning conceptual value, practical value, and our use of quasi-mentalistic language. We’ve offered a systematic way to examine: i) what is violated when human autonomy is hindered or depleted, and ii) what constitutes a well-functioning assistance or augmentation. These are tractable questions for the existing tools of the philosophy of action; answering them is work we hope others will take up.

Acknowledgements

We thank Alex Grzankowski for useful discussions and helpful feedback on an earlier version of this paper. LLMs were used only for polishing and minor edits of author-written text; all edits were author-reviewed.

References

- Ashkinaze, J.; Shen, H.; Avula, S.; Gilbert, E.; and Budak, C. 2025. Deep Value Benchmark: Measuring Whether Models Generalize Deep Values or Shallow Preferences. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Association for Computing Machinery. 2018. *ACM Code of Ethics and Professional Conduct*. Association for Computing Machinery. Accessed 21 May 2026.
- Bales, A. 2025. AI Takeover and Human Disempowerment. *The Philosophical Quarterly*, 75(3): 791–811.
- Bales, A.; D’Alessandro, W.; and Kirk-Giannini, C. D. 2024. Artificial Intelligence: Arguments for Catastrophic Risk. *Philosophy Compass*, 19(2): e12964.
- Bar-Or Nirman, D.; Weizman, A.; and Azaria, A. 2025. Fool Me, Fool Me: User Attitudes Toward LLM Falsehoods. *IEEE Access*, 13: 200351–200366.
- Bermúdez, J. P.; and Felletti, F. 2021. Introduction: Habitual Action, Automaticity, and Control. *Topoi*, 40(3): 587–595.
- Bishop, J. C. 1989. *Natural Agency: An Essay on the Causal Theory of Action*. Cambridge Studies in Philosophy. Cambridge: Cambridge University Press. ISBN 9780521374309.
- Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.
- Bratman, M. E. 1987. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press. ISBN 9780674458185.
- Bratman, M. E. 2007. *Structures of Agency: Essays*. New York: Oxford University Press.
- Bratman, M. E. 2014. *Shared Agency: A Planning Theory of Acting Together*. New York: Oxford University Press.
- Brown, É. 2026. Recommended Selves: Authenticity and Algorithmic Filtering. *Journal of the American Philosophical Association*, 12(1): 15–34.
- Brynjolfsson, E. 2022. The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence. *Daedalus*, 151(2): 272–287.
- Buchak, L. 2013. *Risk and Rationality*. Oxford: Oxford University Press. ISBN 9780198801283.
- Callard, A. 2016. Proleptic Reasons. In Shafer-Landau, R., ed., *Oxford Studies in Metaethics*, volume 11, 129–154. Oxford University Press.
- Callard, A. 2018. *Aspiration: The Agency of Becoming*. New York: Oxford University Press.
- Cappelen, H.; Goldstein, S.; and Hawthorne, J. forthcoming. AI Survival Stories: A Taxonomic Analysis of AI Existential Risk. *Philosophy of AI*.
- Chalmers, D. J. 2026. What We Talk to When We Talk to Language Models. Manuscript. PhilArchive version 2.
- Chernev, A.; Böckenholt, U.; and Goodman, J. 2015. Choice Overload: A Conceptual Review and Meta-Analysis. *Journal of Consumer Psychology*, 25(2): 333–358.
- Christiano, P. 2019. What Failure Looks Like. AI Alignment Forum.
- Christman, J. P. 2009. *The Politics of Persons: Individual Autonomy and Socio-Historical Selves*. New York: Cambridge University Press. ISBN 9780521760560.
- Clark, A. 2025. Extending Minds with Generative AI. *Nature Communications*, 16: 4627.
- Coeckelbergh, M. 2026. AI and Epistemic Agency: How AI Influences Belief Revision and Its Normative Implications. *Social Epistemology*, 40(1): 59–71.
- Davidson, D. 1963. Actions, Reasons, and Causes. *The Journal of Philosophy*, 60(23): 685–700.
- Davidson, D. 1971. Agency. In Binkley, R. W.; Bronaugh, R. N.; and Marras, A., eds., *Agent, Action, and Reason*, 3–25. Toronto: University of Toronto Press.
- Douskos, C. 2017. Habit and Intention. *Philosophia*, 45(3): 1129–1148.
- Dretske, F. I. 1988. *Explaining Behavior: Reasons in a World of Causes*. Representation and Mind. Cambridge, MA: MIT Press. ISBN 9780262040945.
- Dung, L. 2025. The argument for near-term human disempowerment through AI. *AI & Society*, 40(3): 1195–1208.
- Edwards, B. 2025. Annoyed ChatGPT Users Complain About Bot’s Relentlessly Positive Tone. *Ars Technica*.
- Elster, J. 2000. *Ulysses Unbound: Studies in Rationality, Precommitment, and Constraints*. Cambridge: Cambridge University Press.
- Enç, B. 2003. *How We Act: Causes, Reasons, and Intentions*. New York: Oxford University Press. ISBN 9780199256020.
- European Commission. 2020. White Paper on Artificial Intelligence: A European Approach to Excellence and Trust. Technical Report COM(2020) 65 final, European Commission, Brussels.
- Fischer, J. M.; and Ravizza, M. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.
- Frankfurt, H. G. 1971. Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, 68(1): 5–20.
- Gabriel, I. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3): 411–437.
- Gilbert, M. 1987. Modelling Collective Belief. *Synthese*, 73(1): 185–204.
- Goldman, A. I. 1970. *A Theory of Human Action*. Englewood Cliffs, NJ: Prentice-Hall.
- Greenblatt, R.; Denison, C.; Wright, B.; Roger, F.; MacDiarmid, M.; Marks, S.; Treutlein, J.; Belonax, T.; Chen, J.; Duvenaud, D.; Khan, A.; Michael, J.; Mindermann, S.; Perez, E.; Petrini, L.; Uesato, J.; Kaplan, J.; Shlegeris, B.; Bowman, S. R.; and Hubinger, E. 2024. Alignment faking in large language models. arXiv:2412.14093.
- Grice, H. P. 1975. Logic and Conversation. In Cole, P.; and Morgan, J. L., eds., *Syntax and Semantics, Volume 3: Speech Acts*, 41–58. New York: Academic Press.
- Gunn, H.; and Lynch, M. P. 2021. The Internet and Epistemic Agency. In Lackey, J., ed., *Applied Epistemology*, 389–409. Oxford University Press.

- Hartley, K.; Hayak, M.; and Ko, U. H. 2024. Artificial Intelligence Supporting Independent Student Learning: An Evaluative Case Study of ChatGPT and Learning to Code. *Education Sciences*, 14(2).
- Hernández-Orallo, J.; and Vold, K. 2019. AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, 507–513. New York, NY, USA: Association for Computing Machinery.
- Herrmann, D. A.; and Levinstein, B. A. 2025. Standards for Belief Representations in LLMs. *Minds and Machines*, 35(1): 1–25.
- Holton, R. 2009. *Willing, Wanting, Waiting*. Oxford: Oxford University Press.
- Hyman, J. 2017. *Action, Knowledge, and Will*. Oxford University Press, reprint edition. ISBN 9780198769316.
- Iyengar, S. S.; and Lepper, M. R. 2000. When Choice Is Demotivating: Can One Desire Too Much of a Good Thing? *Journal of Personality and Social Psychology*, 79(6): 995–1006.
- Jha, T.; Everitt, T.; and Grzankowski, A. 2026. Human Amplification, Intelligent Agents, and the Aims of AI Research. In *Second Conference of the International Association for Safe and Ethical Artificial Intelligence (IASEAI'26)*. Conference paper.
- Korsgaard, C. M. 1996. *The Sources of Normativity*. Cambridge: Cambridge University Press. ISBN 9780521559607.
- Kosmyna, N.; Hauptmann, E.; Yuan, Y. T.; Situ, J.; Liao, X.-H.; Beresnitzky, A. V.; Braunstein, I.; and Maes, P. 2025. Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task. arXiv:2506.08872.
- Kulveit, J.; Douglas, R.; Ammann, N.; Turan, D.; Krueger, D.; and Duvenaud, D. 2025. Position: Humanity Faces Existential Risk from Gradual Disempowerment. In *Forty-second International Conference on Machine Learning Position Paper Track*.
- Laitinen, A.; and Sahlgren, O. 2021. AI Systems and Respect for Human Autonomy. *Frontiers in Artificial Intelligence*, 4: 705164.
- List, C.; and Pettit, P. 2011. *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford: Oxford University Press.
- London, A. J.; and Heidari, H. 2024. Beneficent Intelligence: A Capability Approach to Modeling Benefit, Assistance, and Associated Moral Failures Through AI Systems. *Minds and Machines*, 34(4): 41.
- Lynch, M. P. 2016. *The Internet of Us: Knowing More and Understanding Less in the Age of Big Data*. New York: Livwright.
- MacIntyre, A. C. 1982. *After Virtue: A Study in Moral Theory, 2007 Edition*. University of Notre Dame Press, 3rd edition. ISBN 9780268035044.
- Mele, A. R. 1992. *Springs of Action: Understanding Intentional Behavior*. New York: Oxford University Press, 1st edition. ISBN 9780195071146.
- Mele, A. R. 2003. *Motivation and Agency*. New York: Oxford University Press. ISBN 9780195156171.
- Munn, L. 2023. The Uselessness of AI Ethics. *AI and Ethics*, 3: 869–877.
- Murray, R. L.; Leonardi-Bee, J.; Barker, A.; Brown, O.; and Langley, T. 2022. The Effect of Alcohol Marketing on People with, or at Risk of, an Alcohol Problem: A Rapid Literature Review. Technical report, University of Nottingham. Report funded by Alcohol Focus Scotland.
- Nelson, M. 2024. Propositional Attitude Reports. In Zalta, E. N.; and Nodelman, U., eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2024 edition.
- Nguyen, C. T. 2020. Echo Chambers and Epistemic Bubbles. *Episteme*, 17(2): 141–161.
- Nogge, R. 1996. Manipulative Actions: A Conceptual and Moral Analysis. *American Philosophical Quarterly*, 33(1): 43–55.
- Nozick, R. 1974. *Anarchy, State, and Utopia*. New York: Basic Books.
- OECD. 2024. OECD AI Principles Overview. Updated May 2024. Accessed 21 May 2026.
- Ouellette, J. A.; and Wood, W. 1998. Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological Bulletin*, 124(1): 54–74.
- Pollard, B. 2006. Explaining Actions with Habits. *American Philosophical Quarterly*, 43(1): 57–69.
- Pollard, B. 2008. *Habits in Action: A Corrective to the Neglect of Habits in Contemporary Philosophy of Action*. Saarbrücken: VDM Verlag Dr. Müller. ISBN 9783836469821.
- Prunkl, C. 2024. Human Autonomy at Risk? An Analysis of the Challenges From AI. *Minds and Machines*, 34(3): 1–21.
- Russell, S. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.
- Schechtman, M. 1996. *The Constitution of Selves*. Ithaca, NY: Cornell University Press. ISBN 9780801431678.
- Scheibehenne, B.; Greifeneder, R.; and Todd, P. M. 2010. Can There Ever Be Too Many Options? A Meta-Analytic Review of Choice Overload. *Journal of Consumer Research*, 37(3): 409–425.
- Schlosser, M. 2019. Agency. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2019 edition.
- Schwitzgebel, E. 2023. How We Will Decide that Large Language Models Have Beliefs. *The Splintered Mind*. Accessed 22 May 2026.
- Searle, J. R. 1990. Collective Intentions and Actions. In Cohen, P. R.; Morgan, J.; and Pollack, M. E., eds., *Intentions in Communication*, 401–416. Cambridge, MA: MIT Press.

- Seidman, J. 2009. Caring and the Boundary-Driven Structure of Practical Deliberation. *Journal of Ethics and Social Philosophy*, 3(1): 1–37.
- Shepherd, J. 2021. Intentional Action. In *The Shape of Agency: Control, Action, Skill, Knowledge*, 57–91. Oxford: Oxford University Press.
- Sher, G. 2009. *Who Knew? Responsibility Without Awareness*. New York: Oxford University Press.
- Strawson, G. 2004. Against Narrativity. *Ratio*, 17(4): 428–452.
- Stray, J.; Halevy, A.; Assar, P.; Hadfield-Menell, D.; Boutilier, C.; Ashar, A.; Bakalar, C.; Beattie, L.; Ekstrand, M.; Leibowicz, C.; Moon Sehat, C.; Johansen, S.; Kerlin, L.; Vickrey, D.; Singh, S.; Vrijenhoek, S.; Zhang, A.; Andrus, M.; Helberger, N.; Proutskova, P.; Mitra, T.; and Vasan, N. 2024. Building Human Values into Recommender Systems: An Interdisciplinary Synthesis. *ACM Trans. Recomm. Syst.*, 2(3).
- Susser, D.; Roessler, B.; and Nissenbaum, H. 2019. Technology, Autonomy, and Manipulation. *Internet Policy Review*, 8(2).
- Suzgun, M.; Gur, T.; Bianchi, F.; Ho, D. E.; Icard, T.; Jurafsky, D.; and Zou, J. 2025. Language models cannot reliably distinguish belief from knowledge and fact. *Nature Machine Intelligence*, 7: 1780–1790.
- Taylor, C. 1989. *Sources of the Self: The Making of the Modern Identity*. Harvard University Press. ISBN 9780674824263.
- Tuomela, R.; and Miller, K. 1988. We-Intentions. *Philosophical Studies*, 53(3): 367–389.
- Université de Montréal. 2018. Montréal Declaration for a Responsible Development of Artificial Intelligence.
- Velleman, J. D. 1992a. Against the Right to Die. *The Journal of Medicine and Philosophy*, 17(6): 665–681.
- Velleman, J. D. 1992b. What Happens When Someone Acts? *Mind*, 101(403): 461–481.
- Velleman, J. D. 1997. How to Share an Intention. *Philosophy and Phenomenological Research*, 57(1): 29–50.
- Watson, G. 1975. Free Agency. *The Journal of Philosophy*, 72(8): 205–220.
- Watzl, S. 2017. *Structuring Mind: The Nature of Attention and How It Shapes Consciousness*. Oxford: Oxford University Press.
- Williams, B. 1981. Internal and External Reasons. In *Moral Luck: Philosophical Papers 1973–1980*, 101–113. Cambridge: Cambridge University Press.
- Williawa. 2025. This is ridiculous. The only memories / custom instructions I have are that I "like pickle rick references", "like mathematical equations" ... X post. Posted April 28, 2025. Accessed May 21, 2026.
- Woolf, V. 1929. *A Room of One's Own*. Norderstedt: Wisehouse Classics. ISBN 9789176376010. 2020 Edition.
- Wu, W. 2013. Mental Action and the Threat of Automaticity. In Clark, A.; Kiverstein, J.; and Vierkant, T., eds., *Decomposing the Will*, 244–261. Oxford University Press.
- Wu, W. 2014. *Attention*. London: Routledge.