Intention Reconsideration in Artificial Agents: a Structured Account

Fabrizio Cariani

[forthcoming in a special issue of *Philosophical Studies* on Normative Theory and AI]

Nicky: Miriam, as friends of yours and Susie's... I'd say more like family than friends at this point. We're just looking out for your well-being. You should consider changing your mind. And then you should change your mind.

from *The Marvelous Mrs. Maisel*, *How you get to Carnegie Hall*

Abstract

An important module in the Belief-Desire-Intention architecture for artificial agents (which builds on Michael Bratman's work in the philosophy of action) focuses on the task of intention reconsideration. The theoretical task is to formulate principles governing when an agent ought to undo a prior committed intention and reopen deliberation. Extant proposals for such a principle, if sufficiently detailed, are either too task-specific or too computationally demanding. I propose that an agent ought to reconsider an intention whenever some incompatible prospect is sufficiently valuable along some dimension that can be assessed at zero or near-zero computational cost.

For a fleeting but powerful moment in the 1980s and 1990s, the philosophy of action and artificial intelligence research found an important point of connection. In "Plans and resource-bounded practical reasoning" (1988), Michael Bratman, David Israel and Martha Pollack outline a joint vision for a modular architecture for a planning agent—an artificial agent able to create, develop and execute plans without specific assumptions about the particular tasks the agent is put to. At the core of this proposal, is Bratman's (1987) theory of intention and practical reasoning, which, they argued, could take up a second job as a blueprint for an artificial agent's practical reasoning module.

In this paper, I argue for the continuing importance of this interdisciplinary effort, for the urgency of theoretical work on one of its main components, and for a specific way of approaching that theoretical work. My focus is going to be on policies of *intention reconsideration* for artificial agents. These are the principles that govern how an artificial agent that is committed to some prior plan ought to be open to reconsidering part or all of it. A substantial body of literature has followed Bratman *et al.* (1988), and it will need to be placed into focus so that my positive goals can be presented more clearly. These are, first, to showcase an important way in which philosophical topics can get new life by being connected to developments in AI, as different standards and considerations may affect theoretical evaluation. And, second, to argue for some programmatic ideas about how to address parts of the unfinished agenda of Bratman *et al.* (1988).

Before approaching these themes, however, I will introduce some relevant background, starting with the general contours of Bratman's account of practical reasoning (§1), then transitioning to its AI application (§2). I then (§3) explain the core constraints on how intention reconsideration is modeled within that AI architecture. The key idea of Bratman *et al.* (1988) is that, while options that are incompatible with the agent's current plans are filtered out by default, some options can override that default. A module in the architecture, the *filter override*, is tasked with identifying these options. After reviewing early attempts to analyze the filter override (§4), I proceed to consider some new proposals. First, I sketch and critique an account that is based on comparisons of desirability between prospects (§5). Next, I critique accounts that classify intention reconsideration as a kind of metareasoning (§6). In the final sections, I describe (§7), give formal

[0]For comments and exchanges on this topic, I am grateful to Boris Babic, Marcello Di Bello, Michael Bratman, Jeff Horty, Thomas Icard, Todd Kahru, Eric Pacuit, the UMD Work in Progress series, the UMD Logic Group. I am particularly grateful to Ilaria Canavotto for written comments on an earlier draft, as well as two reviewers for *Philosophical Studies* for their attentive feedback.

representation to (§8), and finally apply (§9) an account that grounds intention reconsiderations on comparison between individual "determinants" of the agent's preferences.

1 Background: summary of Bratman's approach

Bratman (1987, 1992, 2018) urged grounding the theory of practical reasoning on a representation of a decision-maker as having three types of fundamental attitudes: beliefs, desires, and (future-directed) intentions. Intentions are the most distinctive components of this picture. Very roughly, these are states with the force of commitments to act in certain ways — possibly at a specified time — in response to some decision problem. For illustration, imagine an agent whose goal is to feed a family of four for a week-end. The agent's plan is structured into a bundle of intentions—to secure enough food at the grocery store, to cook a pasta for Saturday's lunch, to prepare a salad for Sunday, and so on.[1]

Not any bundle of intentions makes a plan, however. The intentions in a plan must form a coherent structure, in which individual intentions are linked to others, for example as means to ends. Intending to cook the pasta on Saturday requires me to have the ingredients available at the appropriate time, which is facilitated by an intention to secure the ingredients in the first place. Any committed intention pressures the agent to adopt the means to its realization. So, the intention to secure the ingredients might in turn lead the agent to new deliberation, the outcome of which may itself be a new intention to go to the grocery store. Intentions are formed, maintained and updated, within an environment that is changing in a variety of ways. There could be external changes (e.g., some food might go out of stock); there could be changes in the agent's epistemic access (e.g., the agent might come to learn of the availability of an ingredient that had been presumed to be unavailable); and there could be changes in what the agent desires.

It is central to the account that plans are partial in at least two ways. The first dimension of partiality is that some details that must pertain to any execution of a plan do not belong to the plan itself. The agent might intend to purchase cheese for the pasta, but they won't plan to buy some specific number of grams of

[1] The philosophical literature on future-directed intentions is well developed. For some important representatives, in addition to the aforementioned works by Bratman, see also Harman (1976, 1986, ch.8), Holton (2009, ch.1), Tenenbaum (2018, 2020). Holton helpfully discusses psychological evidence that agents have states with the broad features of intentions (as opposed to merely arguments that it would be rationally valuable to have intentions). For an introductory essay, see chapter 5 of Paul (2020).

it. Crucially, plans are *temporally* partial in that they may specify that some actions are to be undertaken without specifying when. To extend our example, the food must be secured before it is cooked, but there is no specific order in which the pasta or the oil must be added to one's grocery cart. In other contexts, committed actions might lack precise temporal scheduling even if there are constraints on their relative orders.[2]

Selecting appropriate intentions from a menu of options is the main function of deliberation. Deliberation results in committed intentions. However, intentions are not just outputs in the deliberation process: committed intentions play a role in shaping deliberation, by constraining the range of options that are entertained in future rounds of deliberation. Returning to our guiding example, after having settled on the cooking plan, our agent is in a position to ignore irrelevant options, such as visiting music stores (at least as far as the specific plan is concerned). In Bratman's slogan, *plans guide and focus deliberation* (1987, §3.5), by restricting deliberation to a manageably small domain of relevant options. Both the partiality of intentions and their role in focusing deliberation fit well with an important theme in Bratman *et al.* (1988), namely that a theory of intention needs to respect the fact that agents are *resource-bounded*.[3,4]

One of the key functions of intentions is to allow agents to steer a middle course between, on the one hand, the stability that is required to accomplish complex goals, and, on the other, the flexibility that is required to deal with a constantly changing environment. The resiliency of rational intention gets much of the attention in most presentations of this topic, but the flexibility aspect is equally important. Agents have that flexibility because intentions may, in appropriate circumstances, be reconsidered, and ultimately revised. Imagine that you plan to take a trip to New York to see a Broadway musical. You have purchased train tickets, and booked a hotel. At that point, it is announced that one of your favorite artists is playing a show in the DC area. Depending on your priorities, you might consider reopening deliberation and potentially end up revising your plans. In this example, the emergence of a new option prompted you to reconsider your settled

[2] Weld (1994) amplifies the theme of temporal partiality in the AI literature.

[3] In practice, intentions cannot focus deliberation quite as sharply as suggested by Bratman *et al.*, agents will often entertain options that do not pertain to any imminent deliberative needs in the context of conditional deliberation. An agent might plan a bike route for a trip they will take if it is not raining over the weekend, and also plan a museum trip they will take if it is raining. Even in conditional deliberation, intention can play this role of constraining the range of hypotheses that get entertained.

[4] The theme of boundedness and agency is rich both in philosophy (see e.g. Millgram, 2019), economics (see Conlisk, 1996, for an overview), and obviously in AI research (see, e.g. Russell, 2016).

commitments. Of course, intentions may also be reconsidered for reasons other than the emergence of new options, such as acquiring new beliefs (e.g. learning that the musical was postponed) or changing one's desires (e.g. ceasing to enjoy attending Broadway musicals).

Two related theoretical questions are raised by the possibility of reconsideration, and it is critical to keep them distinct. First up, when should an agent reconsider some settled intention? And assuming that some new alternative is chosen once deliberation is re-opened, how should revision proceed? Call the first the *reconsideration* question and the second the *revision* question. Reconsideration is a matter of undoing a commitment to an intention, thus reopening deliberation on the relevant question; if the new deliberation results in a change of heart, the agent will have to perform revision to secure the internal coherence of their resulting plans. (E.g. if, on reflection, they decide to attend the concert in DC instead of the Broadway Musical in NY they ought to revise their intention to book a hotel in NY.) As the present discussion transitions from the theory of rational practical reasoning to applications to artificial agents, these questions also assume an algorithmic and formal dimension. Contributions like van der Hoek *et al.* (2007); Lorini and Herzig (2008), and Icard *et al.* (2010) address the second question from the point of view of an extended analogy between formal systems of belief revision and formal systems of intention revision. My focus here will be on discussions of the reconsideration question in the AI literature, and primarily within the broad family of views that emerge from Bratman's philosophy of action.

2 Background: the IRMA planning architecture

Bratman's account of intention fueled a research program in AI devoted to the development and analysis of the BDI (=*Belief*, *Desire*, *Intention*) framework—a framework for artificial practical reasoning. Before introducing any more details, it is worth pausing and appreciating that this is a notable instance of a philosophical theory having significant impact within a field of AI.[5] The intellectual context in which this AI/philosophy of action link emerged was a gradual coming into focus of the idea of developing an *intelligent agent* as a central theoretical target for AI research. (Russell and Norvig 1995, ch. 1; Thomason and Horty 2022,

[5] At the time of this writing, Bratman *et al.* (1988) has been cited two thousand times. For a survey of contributions that explore the implications of IRMA for the philosophy of action and agency, see Thomason and Horty (2022). For more general overviews of the impact of the BDI model in AI, see Georgeff *et al.* (1998), Wooldridge (2009). For a recent discussion of the logical legacy of, and open challenges within, the BDI framework, see Herzig *et al.* (2017). Bratman's influence is noted also in the currently standard AI textbook (Russell and Norvig, 1995, p.1041).

p.366). What is meant by 'intelligent agent' here is an entity that (i) is located in a dynamically changing environment; (ii) is endowed with a goals of many kinds; (iii) is capable of learning from the environment; and (iv) is, crucially, capable to act on it in ways that further its goals. A philosophical theory of instrumental rationality (such as Bratman's), may count as a starting point, and candidate proposal, for building an intelligent agent. Conversely, philosophical research on instrumental rationality might benefit from reflection on how such an agent ought to be built. It goes without saying that neither theoretical effort should simply defer to the other, and they are to be pursued in parallel.

Conceiving of the problems of instrumental rationality from the perspective of this sort of AI application comes with certain conceptual advantages. According to a prominent line of criticism of Bratman's account of practical rationality, intentions are not fundamental mental states. According to this objection, while it may be granted that individual intentions are not identical to any particular desires, it is nonetheless the case the intentions may be reduced to, and indeed constituted by, complexes of beliefs and desires (Velleman, 1989; Ridge, 1998). It seems to me, however, that from the point of view of the AI application, the question of whether intentions are reducible to combinations of beliefs and desires is of little or no importance.[6] It is consistent with the spirit of the design of an artificial agent that there could be important theoretical reasons to single intentions as having a special role within the architecture that are independent of matters of reduction and internal constitution. Since it is plausible that "mere" desires and intentions play different roles in the architecture, there is enough justification for singling out intentions as special.

The specific BDI architecture proposed by Bratman *et al.* (1988) has come to be known as IRMA (*Intelligent Resource-bounded Machine Architecture*).[7] The architecture is based on four data structures: beliefs, desires, intentions structured into plans, and a plan-library (a kind of background catalog of all possible plans the agent knows about). In its central loop, an IRMA agent initializes, fleshes out, and stores a (typically partial) plan. The main process in this loop is deliberation, whose role is, as in Bratman's account of practical reasoning, to add new intentions

[6] I do not mean to imply here that the philosophical question of the reducibility of intentions is decisively solved in favor of the reductionists. For a critique of reductionist approaches, see (Holton, 2009, pp.17-19). Even earlier works, such as (Harman, 1986, ch.8) already strongly suggest that regardless of the question of reducibility we get our best grip on intentions by reflecting on how they function in the context of agency.

[7] As far as I can see, this label does not occur in the 1988 article itself. It is however referenced as such as early as 1990 (Pollack and Ringuette, 1990).

to plans.[8] New intentions are chosen from a menu of *options* on the basis of beliefs, desires, and the previously settled intentions. Figure 1 replicates, with simplifying omissions, IRMA's diagrammatic representation in Bratman *et al.* (1988).

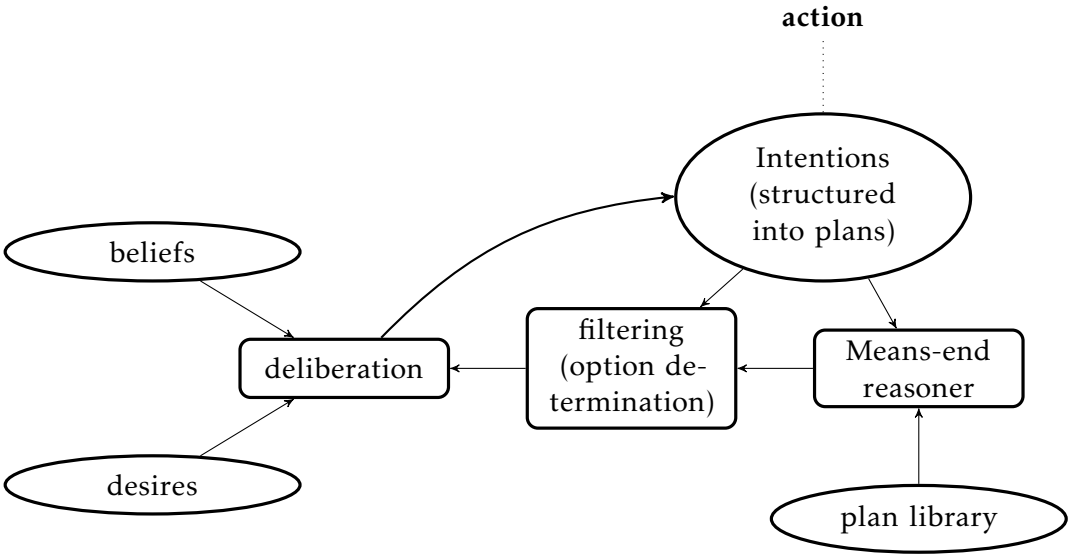

Figure 1: A simplification of Bratman *et al.* (1988)'s diagram of IRMA

The architecture itself is emphatically neutral on the internal shape of deliberation. In the interest of determinacy for my later discussion, I am going to assume that deliberation is powered by some kind of expected value calculation, with quantitative analogues of beliefs and desires being recruited in attaching expected values to options. Intentions play a role in deliberation by influencing the process that identifies options. (See Bratman 1992, p.4 on "framework reasons".) As noted, they do so because options that are incompatible with the established intentions in the plan get filtered out, and thus ignored, by default.

However, it is crucial to IRMA's ability to capture the possibility of intention reconsideration that that default can be overridden. The effect of overriding it is to submit to deliberation an option that would otherwise have been ignored. Importantly, overriding the default does not automatically incorporate the option into the plan. In the terminology I established in the previous section: reconsideration does not automatically entail revision. The consequence of a successful override is that an otherwise excluded candidate option gets restored to the roster

[8] Needless to say, the idea of a loop for planning centered around deliberation is not specific to the BDI framework. It plays a central organizing role in one of the leading textbooks on automated planning. In particular, Ghallab *et al.* (2004) is organized about different models of automated deliberation, suitable to a variety of planning circumstances.

of available options in deliberation. Consider the case of my planned Broadway musical trip, as described in the previous section. Once the competing concert in DC is announced, the filter override module suspends the default elimination of incompatible options. As a result of this suspension, some options that are incompatible with a trip to Broadway get incorporated into deliberation, at which point the role of the filter override mechanism is complete. If the new option is adopted, it will prompt revisions across the board—travel reservations must be cancelled, tickets must be sold on the used market, and so on.

3 Core constraints on filter override

Because the point of these mechanisms is to constrain deliberation so as to make it feasible for resource-bounded agents, the operation of the filter override must be computationally cheaper than deliberation itself (Bratman *et al.*, 1988; Bratman, 1992; Pollack and Ringuette, 1990; Schut *et al.*, 2004). There would be no point to a reconsideration mechanism if the resources that are required to deploy it are as costly as what's required in deliberation. If there was no computational disadvantage to doing so, one could just deliberate from scratch among all the available options all the time.

This observation is significant on a few different levels. In applications in which computational capacity is effectively unbounded—or unbounded relative to the task—there is no need for a dual-layer architecture such as IRMA. As Pollack and Ringuette (1990, p. 186) put the point, "if deliberation is extremely simple, it may be redundant to posit separate deliberation and filtering processes". This might undermine the interest of the project if one thought that computational power was effectively unbounded for the kinds of planning problems that AI researchers and engineers care about. In my view, however, the idea that the problems of interest are exclusively the ones in which deliberation is computationally cheap comes from a narrow view of what we ought to demand an artificial agent with a full capacity for planning. In many highly constrained environments, we may rely on brute computational capacity to drive intention reconsideration. But it is evident that this is not generally the case for many kinds of ordinary plans, which usually include many examples of plan reconsiderations that are potentially open-ended and unconstrained. To use an example with a distinguished history in artificial intelligence, choosing the topic of a conversation, and choosing when to change the conversation away from an established topic, is clearly part of planning behavior, but very much not something where we can rely on brute computational

capacity to drive the dynamics of speaker intention. Another example could be planning for completion of a partially developed work of art or script. As we move our attention to creative planning, the idea that deliberation is guaranteed to be computationally cheap ought to recede to the background.

Another implication of the gap between filter override and deliberation is that the filter override must be at best an imperfect predictor of the outcome of deliberation. A *perfectly predictive* override module would have this feature: whenever an option is restored among the options in deliberation by the override module, it is also selected in deliberation. In saying that the filter override cannot be perfectly predictive, I do not mean that this would be undesirable. The point is that a perfectly predictive override module would obviate the need for a separate, computationally more intensive deliberation module. In this respect the structure of IRMA is vaguely reminiscent of dual process theories in cognitive psychology, according to which human cognition can be modeled by distinguishing two systems — roughly speaking one that is heuristic and fast, and one that is analytical and deliberate.[9]

4 Early analyses of intention reconsideration

In light of the goal of developing an architecture, Bratman *et al.* (1988) do not provide specifics about the content and internal structure of the filter override. Though this decision fits with the goals of their paper, it tasks subsequent literature with fleshing out the details of this important module. With some important exceptions, however, the subsequent literature has either been silent on this matter or has replaced the problem of spelling out the details of the filter override with some slightly different problems. In particular, much work within the BDI tradition but outside of the IRMA framework has distinguished itself in part by renouncing the idea of analyzing intention reconsideration in terms of filter override For example, Wooldridge 2009 explicitly singles out the filter override mechanism for omission in his textbook presentation of BDI agents.

Among the exceptions to this generalization is Pollack and Ringuette's influential experimental analysis of the *Tileworld* environment (Pollack and Ringuette, 1990). In the Tileworld environment a lone agent moves around a square grid with the goal of filling "holes" (see Figure 2, which is lifted directly from Pollack and Ringuette 1990). Some cells on the grid are distinguished by a numerical values

[9]See among many, Sloman (1996), Evans and Stanovich (2013) and the review essay Osman (2004).

and holes are sets of contiguous cells with the same numerical value. That value indicates how many points the agent will obtain if she fills the hole. To fill a hole, an agent must slide tiles to cover each numerical cell in the hole. Once a hole is filled, it disappears, together with the tiles that filled it in the first place. To push tiles, and in general to move around the world, the agent can move either horizontally or vertically along the grid, but not diagonally (Pac-Man style!). The grid features several obstacles — cells that the agent cannot walk through (similarly, they are also impassable for tiles).[10] The agent's quest in the Tileworld environment is not meant to be a game: Pollack and Ringuette note that the Tileworld environment provides an abstract representation of the Robot Delivery Model, in which a robot navigates an office environment to distribute messages of the appropriate kind to workers stationed at their posts. In recognition of this fact I will refer to this as the "Tileworld task".

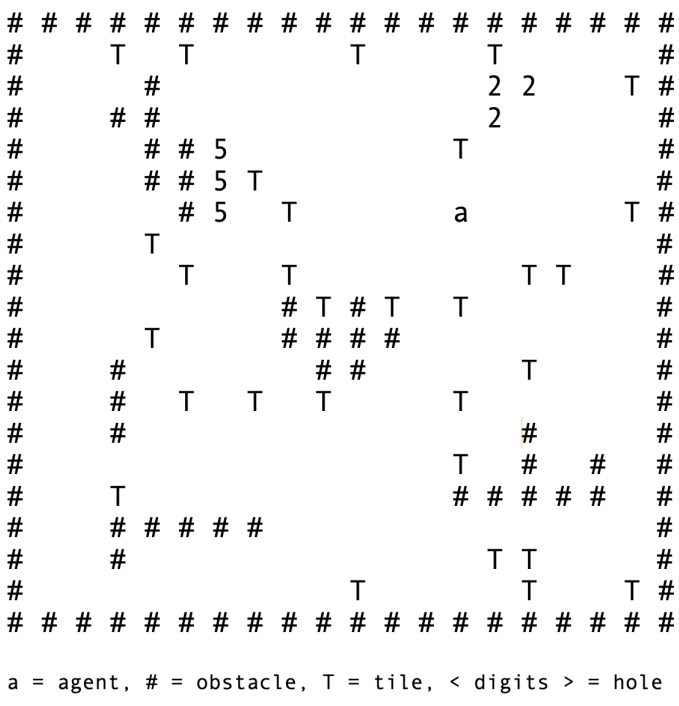

Figure 2: Sample representation of tileworld starting state (replicated from Pollack and Ringuette 1990).

[10]The original Tileworld setup introduces complexity that is not needed to convey many of the central insights that can arise from its study. Many later studies stripped off elements such as e.g., removing the tiles (and instead allowing the agent to fill a hole just by walking to it), removing obstacles, turning holes from pluralities of cells to individual cells, and many others (Kinny and Georgeff, 1991; Schut *et al.*, 2004).

The actions of the agent in Pollack and Ringuette's experimental study are guided and organized by an implementation of the IRMA architecture. The agent deliberates as to which hole to fill and by means of which tile(s), then sets it as its intention to fill that hole with that particular tile (or group of tiles). In assessing whether to fill a hole, some properties of the hole are clearly relevant: its points value, its distance from the agent, and its distance from the closest tiles.

Pollack and Ringuette note two possible implementations for a deliberation module. The simpler one is just to choose to fill whichever hole has the highest score. The more sophisticated one goes after a kind of estimate of a utility calculation for the agent aiming to complete the task. The exact formula is based on what they call a 'likely value', LV of a hole, where, provided that h contains n tiles and there are n or more tiles available, we let:

$$LV(h) = \frac{score(h)}{dist(a,h) + \sum_{i=1}^{n} 2 * dist(h,t_i)}$$

Informally, the likely value of a hole is an aggregate of the score of the hole, of the distance from the agent to the hole, and from the hole to the tiles. As Pollack and Ringuette note, this is just one example of a deliberation module for a Tileworld agent.[11] Executing the intention requires that the agent move towards the hole, and at each tick of the clock new holes might respawn in different locations in the world. As they do, the agent might need to decide whether to go after those new opportunities or stick with the one it is targeting. In order words, it needs to be able to reconsider its intentions.

In Pollack and Ringuette's implementation of the IRMA architecture, the filter override is analyzed as follows. Let v be parameter with integer values. Then, filter override is triggered iff:

$$(\texttt{alt goal points} - \texttt{current goal points}) > v$$

The dependence on a parameter allows them to model agents with different dispositions towards the possibility of reconsideration. Indeed, Pollack and Ringuette even leave open the possibility of negative values for v. In the extreme case, $-\infty$ is also allowed, representing the maximally permissive filter, which corresponds to the "cautious" strategy of filtering nothing and submitting every option to deliberation.

[11] Note that this is not an *expected* utility estimate. This is because there is minimal uncertainty in the Tileworld task. Indeed, the only elements of uncertainty for the agent have to do with where and when new holes will spawn. However, since this is completely random it would seem that this uncertainty has no clear relevance to the agent's deliberation. As a result, under the standard stipulations, there is no reason to consider expectations.

The glaring theoretical limita of this analysis is that it is too tied to the specifics of the Tileworld environment. Naturally, a simplified approach is wholly unproblematic for the purposes of running simulations and comparing various strategies to the Tileworld task. But a theory of intention reconsideration for AI agents needs broader scope—or, at least, a general formulation from which task-specific accounts can be derived as instances. The point comparison account might need significant changes even in small variants of the Tileworld task—for example, if the agent's aim is to hit 100 points in the shortest possible time.

Kinny and Georgeff (1991) also produce experimental analyses of various approaches to intention reconsideration in a variant of the Tileworld scenario. Their account of intention reconsideration does not lean on a task-specific theory of intention reconsideration. However, they also let the modeling needs dictate an analysis that seems too basic and coarse-grained for our purposes. Instead of the filter override mechanism, they consider *time-based* reconsideration policies. After a fixed number of rounds, the agent reconsiders their plans by deliberating from scratch, based on newly available options. Specifically, Kinny and Georgeff consider three types of agents: a **cautious** agent (who reconsiders their intentions at each tick of the clock); a **middling** agent (who reconsiders after some number $k > 1$ of ticks); and a **bold** agent (who never reconsiders until the intention that is currently in focus is complete). Before moving to more substantive observations, I will go on the record as deprecating the (unfortunately established) use of "bold" in this terminology. Boldness is not a particularly salient trait of agents who never reconsider their plans. In fact, one might think that boldness is associated with *changing* one's plans. I refer to such agents as **stubborn**.

By describing Kinny and Georgieff's analytical paradigm as 'too coarse-grained' I do not mean to suggest that it is not of theoretical value. They use simulations to support several important intuitions we might have about reconsideration. For example, they illustrate how in a slow-changing environment stubborn agents perform better (since they do not waste time reopening deliberation); in a very fast-changing environment, it is cautious agents that perform best. Many of the morals of these kinds of experimental studies are robust as more nuanced reconsideration models are considered.

However, it also seems plausible that more substantive policies of intention reconsideration, such as the filter override approach of the IRMA architecture, will be strictly better than time-based approaches—provided that they are based on elements that are more significant towards adapting the agent's goals to her

circumstances. The point of being a cautious agent is not to second-guess every decision every time a new option comes along. Instead, it is to scout the environments for opportunities, and to monitor changes within one's own attitudes that might lead to changes of heart.[12] The filtering approach suggests, plausibly, that whether an agent ought to reconsider their intention depends on how the environment and the agent themselves have changed, and not merely a reflection of the passage of time. A Tileworld agent who is aiming to fill a 7-point hole, need not be distracted with a 2-point hole, but should perhaps give more careful consideration to a 9-point hole.

5 The desirability-based account

Taking a step back, it seems preferable to adopt as a starting point Pollack and Ringuette's account of filter override, and generalize from there. Instead of comparing scores, suppose that the filter override is triggered when there is a filtered-out option whose desirability to the agent is sufficiently larger than the current goal's:

$$(\texttt{alt goal desirability} - \texttt{current goal desirability}) > \tau$$

Here I am thinking of desirability as an abstract concept which registers the strength of a desire and can play an explicit role in deliberation.[13]

Formally and conceptually, the concept of desirability I have in mind shares elements with the decision-theorist's concept of utility, but I intentionally chose not to use the term 'utility'. In standard decision theory, the concept of utility comes with substantial theoretical associations, and it is important to avoid them. Here is Kenny Easwaran highlighting the contrast I have in mind:

> Naive applications of decision theory often assume that it works by taking a specification of probabilities and utilities and using them to calculate the expected utilities of various acts, with a rational agent being required to take whichever act has the highest (or sufficiently high) expected utility. However justifications of the formal framework of expected utility generally work in the opposite way—they start

[12] Within the BDI tradition that eschews the filter override approach, Schut *et al.* (2004, esp §2.4 ff.) discuss this point with clarity and entertain some more nuanced reconsideration policies.

[13] Further elaborations of this idea are possible. For example, one might consider parametrizing the threshold to a time, or to a feature of the agent's psychology or to a deliberative context. These elaborations would reflect the fact that the circumstances and make-up of the agent can impact their propensity to reconsider. The very same agent ought to be stubborn in some contexts and cautious in others. I do not pursue these elaborations here.

> with an agent's preferences among acts, and use them to calculate an implied probability and utility function. (Easwaran, 2014, p.1)

Utility is a representational device that is used to construct a structured representation of an agent's preference ordering. It is not a concept tracking some quantity that is to be used by the agent in deliberation. Relatedly, the aim of decision theory on its standard interpretation is not to provide a recipe that the agent is supposed to follow in deliberation. Even for those who lean towards more 'realist' construals of credence and utility, the point of decision theory is to characterize a type of incoherence between an agent's beliefs, desires, and preferential states. My use of 'desirability' here reflects the need for a more generic term that is free of those particular associations. Those who prefer to stick to the prevailing terminology, might think of 'desirabilities' as short for 'naïve utilities' in Easwaran's sense.

With that digression out of the way, let us move to evaluating the desirability comparison account, starting with its advantages. By relying on a more abstract concept of desirability, the theory allows agents to reconsider intentions in situations in which score alone is not the right metric. Imagine a Tileworld agent who, instead of aiming to maximize their score, is aiming to hit the 100 points threshold as early as possible. The agent sits at 99 points and plans to fill a 5-point hole. At that time, a 1-point hole spawns right near them, though not directly on their current path. The new hole is an opportunity to hit the goal earlier, but score-comparing filters are insensitive to this kind of prospect. Examples like this can also be considered in the original Tileworld setting: even for the agent who is trying to maximize score, it might be advantageous to take a small detour and thread through both the new and the old hole.

A minor problem with the desirability approach is that, as far as desirability measurements are concerned, the numerical magnitudes of differences are not meaningful quantities. I emphasized that I want to keep the concept of desirability distinct from the decision-theoretic concept of utility. But this much they have in common: both can be assumed to be measured by interval scales. A consequence of this is that any scales obtained by multiplying any given desirability scale by a positive linear transformation are equivalent (a positive linear transformation is of the form $ax + b$, for a a positive real number, and b a real number). Concretely, saying that two prospects are 20 desirability points apart is no more meaningful than saying that two rooms are 20 degrees apart in temperature, without specifying the temperature scale. It is easy enough to patch this problem. The intuition behind the desirability account is that the alternate prospect should be sufficiently

more desirable than the current one. Although this cannot be represented by a number regardless of scale, it can be represented by a number relative to a scale. So, where D is the desirability function, and τ_D a threshold that depends on D, consider this generalization for the filter override:

$$(\texttt{alt goal desirability} - \texttt{current goal desirability}) > \tau_D$$

Note that the fact that D is an interval scale means that there is no conceptual point to distinguishing between positive and negative values on the scale associated with the range of the function. (Of course, this does not mean that there is no point to distinguishing between positive and negative values of τ_D.)

Far more serious problems loom for the desirability account. The most significant is that it narrows the gap between filter overriding (i.e. reconsideration) and deliberation, to the point where filtering appears almost exactly as computationally expensive as deliberation. Recall that a constitutive component of the filter override is that it is to function as a computationally cheaper surrogate for full-blooded deliberation. Since deliberation is itself plausibly driven by comparisons of desirability, we seem to have lost sight of the distinction between what warrants reconsideration and what warrants deliberation.

Furthermore, in switching from points to numerical representations of desirability of prospects, we have switched from a quantity which was assumed to be transparent (that is: known at zero cost) to the agent to one that is arguably not. It was part of the stipulation of the Tileworld task that the agent knows at any point the state of the world with regards to hole values and hole distances. By contrast, the overall desirability of any one option cannot in general be assumed to always be known at zero cost by the agent. Plausibly, the overall desirability of a plan in Tileworld is a complex function of the value of the holes, the distance of the hole from the agent, the distance of usable tiles, as well as even more general factors such as the agent's broader goals in the Tileworld task. This is not a specific quirk of the Tileworld environment: there is no easy path from the kinds of things we have immediate access to to the overall desirability of the plan.

6 The metareasoning account

According to one prominent perspective in more recent literature (Schut and Wooldridge, 2000, 2001; Schut *et al.*, 2004; Van Zee and Icard, 2015), intention reconsideration may be viewed as an instance of "metareasoning". The investigation of metareasoning is a major area of reasearch at the interface of AI and Cognitive

Science (see, among others, Russell and Wefald 1991; Fletcher and Carruthers 2012; Russell 2016; Griffiths *et al.* 2019, and references therein).

Simplifying somewhat, in the metareasoning perspective an agent is viewed as engaging in deliberation at multiple levels. The base level is ordinary evaluation of options. This may proceed by comparing expected desirabilities. Another level is evaluation of higher-order questions, such as whether to stick with one's committed options or reopen deliberation. Among the works on metareasoning and intention reconsideration referenced in the prior paragraph, Schut *et al.* (2004) provide the most comprehensive analysis of Tileworld and I focus on their discussion here.

One way of developing the metareasoning approach is to view deliberation as an action in its own right. Given that, the agent must engage in a kind of metadeliberation, concerning whether to deliberate or not. One way to make this concrete is to have the agent compare the option of deliberating against a "default" option (Russell and Wefald, 1991). In the application to Tileworld, this can be assumed to be the agent's committed intention. Of course, any such view must face up to the constraint that the computational cost of reconsideration be meaningfully lower than that of deliberation. One approach to meet this constraint is to have the agent perform ordinary deliberations by comparing expected desirabilities, but metadeliberations by comparing *estimated* expected desirability. For example, the value of deliberation might be estimated to equal the desirability the agent would receive on the assumption that a (fixed) likely sequence of actions is undertaken, or perhaps that deliberation would lead to certain likely outcomes.

Schut *et al.* (2004) develop two architectures for such an agent in a Tileworld setting. In the interest of space, I only discuss the first of them.[14] In it, the agent compares two estimates of desirability:

q_1 **The estimate of the desirability of the agent's committed intention**.
This is implemented as the agent's distance in steps from the "intended hole".

q_2 **The estimate of the desirability of deliberation.**
This is modeled as the ratio `avedist/newholes` where `avedist` is the average distance of the agent from any location on the board and `newholes` is

[14] The second model, which is based on Partially Observable Markov Decision Process, also has plenty of interest. The framework of Partially Observable Decision Processes is the gold standard for the analysis of sequential decision making at the interface between decision theory and AI (see Russell and Norvig 2020, ch. 17, and Icard forthcoming, §2.2-2.4). However, the specific development of the POMDP approach to tileworld generates similar concerns to the ones outlined in the main text with regards to the first approach.

the estimate of number of new holes that spawned since the agent's last deliberation.

If q_2 exceeds q_1, the agent sticks to their committed intention, otherwise they reopen deliberation. Note that, roughly and generally speaking, `avedist` is higher if the agent is located at the edges of the tileworld, and lower if the agent is closer to the center. Thus agents are more likely to stick to their plans if they find themselves at the edges of the board, and if they estimate that relatively few new holes have appeared.

In my view, it is doubtful that these two quantities are good estimates of (expected) desirability. In particular, the two components of q_2 seem to be two instances of a much larger class of properties that could be used to estimate the value of deliberation. I agree that the position of the agent and the amount of new holes do affect the deliberation. But so do the values of the holes, the agent's proximity to any new holes that might have spawned, and so on. Similar considerations also should make us question whether q_1 is a good measure of the desirability of an intended hole. A further point of criticism is that it seems plausible that, in the course of a Tileworld history, the properties that ground the estimate of the desirability of deliberation might change, and thus that no theory at this level of specificity can quite be right. Earlier on I considered an alternative Tileworld task in which the agent's overarching goals is to hit a total of 100 points.

While these concerns target only the specific implementation choices in (the first model of) Schut *et al.* (2004), there is a more fundamental objection on which I would like to rest my argument. The deeper worry is that it is not clear what makes the comparison between q_1 and q_2 an instance of metareasoning at all. In the simplest terms, the agent is running a basic test on their environment. What could possibly make this count as metareasoning?

In an inflationary sense, metareasoning is the result of a dedicated rational module dedicated to metacognitive tasks—such as when to start and stop reasoning and deliberating. In a more deflationary sense, metareasoning is just any kind of monitoring and controlling of "the processes involved in learning and remembering [..] as well as some of the thought processes involved in reasoning and decision making." (Fletcher and Carruthers, 2012, p.1366) In our intended application, deflationary metareasoning occurs whenever agents make a determination between higher-order options, such as whether or not to start/continue/resume reasoning.

It is doubtful that there is anything metacognitive in this inflationary sense in the model of Schut *et al.* (2004). What triggers deliberation in this model is a comparison between two properties that are chosen, somewhat arbitrarily to stand in for estimates of desirability. Given how rough these estimates are, describing this comparison as a higher-order comparison of expected desirabilities adds little or nothing to our understanding. In practice, this reconsideration module could be characterized equally well as requiring the agent to be responsive to some heuristically salient features of their environment.

The model does, of course, count as an instance of metareasoning in the deflationary sense. But deflationary metareasoning is such a broad, heterogeneous category that it is not especially illuminating to characterize any one phenomenon as an instance of it. Crucially, it does not follow that the shape of metadeliberation ought to be in any way like the shape of deliberation. This is because all that deflationary metareasoning requires is any kind of comparison of higher-order options. Indeed, *any* act of reconsideration counts as an instance of metareasoning in the deflationary sense.

One last point: an insight of Bratman's can guide us towards an alternative to the metareasoning paradigm. Specifically, Bratman discusses the rationality of what he calls 'nonreflective (non)reconsideration' . According to Bratman, stability and reconsideration are grounded in:

> various general habits and propensities [...] whose reasonableness we may asses in a broadly consequentialist way. The nonreflective (non) reconsideration of a certain prior intention is rational of *S*'s if it is the manifestation of general habits of reconsiderations that are reasonable of *S* to have (Bratman, 1987, §5.2).

It is clear from the context of this discussion (as well as from late presentations, such as Bratman 2018, pp. 152-153) that in the typical case, reconsideration is more naturally understood as a non-deliberative process. Bratman's main focus are human-like agents, but it provides at least initial motivation to explore non-deliberative models for intention reconsideration in artificial agents. Indeed, the idea could be viewed as a general, implementation-independent insight concerning

the processes that underpin reconsideration.

7 The known determinant account

Let us take stock of where we are and what we need to design. We need a principle for intention reconsideration that is: (i) precise enough that it could be implemented with little extra work, given an implementation of the IRMA architecture; (ii) task-independent; (iii) such that the agent can run the reconsideration module with little or no computational cost; (iv) stays clear of the inflationary conception of metareasoning.

To these ends, I propose that we shift our focus from desirability to its determinants, and in particular to those determinants of desirability that can be known on the cheap.[15] The basic intuition is that the desirability that an agent attaches to each option is not well understood as an atomic lump. Instead, it is always structured by some implicit rubric, and the entries within that rubric are what I will refer to as 'determinants'. Let us say, then, that the determinants of a desirability function are whatever factors inform the values it assigns to options or prospects. The discussion of Tileworld provides some examples of determinants: points attached to holes, distances from the holes and between tiles and holes, and so on.

I want to introduce two different perspectives one might take on this implicit rubric. According to the *direct* perspective, the rubric may be viewed as part of the apparatus that the agent brings to deliberation. The direct perspective is not mandated. According to the *indirect* perspective, we might suppose that the rubric structure doesn't actually play a role in deliberation. Instead, the rubric is generated *post-hoc* as a heuristic guess that models the likely generation of the desirability function. The immediate inspiration for this approach is the literature on explainable AI (XAI). This literature has emerged in response to concerns about the opacity of machine learning algorithms. In that contexts, a class of proposals have been advanced according to which a machine learning algorithm may be coupled with a secondary algorithm whose job is to provide users of the algorithm and stakeholders a rough understanding of the factors that influence the underlying algorithm's classifications. The notable and crucial feature is that this secondary algorithm has no influence on the underlying primary algorithm. The *post-hoc* approach to explainable AI has gone in for serious criticism, as an

[15] For an alternative approach, see Schut *et al.* (2004), who entertain the option of using estimates of desirability.

answer to (admittedly somewhat vaguely formulated) interpretability challenges (Lipton, 2018; Krishnan, 2020; Babic *et al.*, 2021).[16]

But in the present context, I am not suggesting applying the idea in the same context, and it is not clear that the kinds of complaints that are frequently voiced against *post-hoc* explanation methods apply. In the barest terms, the indirect proposal is the view that an agent may maintain at the same time a utility function and a *model* of the utility function. The reconsideration module would be sensitive to the model of the utility function, while deliberation itself reliefs on the utility function.

In the most basic formulations of the Tileworld task, it is plausible to suggest that all the main determinants of desirability are immediately known at zero cost by the agent. In general, however this might fail to be true. Consider a simple variant of the Tileworld environment in which the point-values attached to each hole are only known if the agent is no more than 5 steps away from that hole. (Assume that the agent still maintains zero cost knowledge of the location of the holes.) The scores associated with the holes continue to count as determinants but they cannot be assumed to always be transparent to the agent. This situation is common to everyday human practical reasoning: the desirability I assign to the prospect of enrolling my children in public school, as opposed to private, is embedded in a complicated set of dependencies on attitudes I hold and facts that obtain in my environment.

To keep this distinction in mind I will use the phrase *known determinants* of desirability to refer to those determinants which the agent immediately knows. In practice, the expression is a bit more of a slogan than a hard barrier. In particular, we should not to be strict about what counts as "known" for these purposes. After all, zero cost knowledge is a high standard—one that is rarely met by things we take for granted in real-life instances of planning and reconsidering. It will be best to admit as "known" those determinants for which the agent merely has reliable estimates whose cost is either zero or very low compared to their time pressures.

With that concept in hand, I can state my proposal for the filter override module: filter override should track potential improvements in known determinants of desirability. A filter override is triggered if the new prospect offers to the agent significant improvement along one or more known determinants. To put that in a compact formal statement, say that in a deliberative context C, an option o triggers filter override iff there is a known determinant d_i in the desirability function D

[16] For an interesting defense of the post-hoc approach in the XAI context, see Fleisher (2022).

such that the alternative is significantly better than the current intention:

$$d_i(\texttt{alternative}) - d_i(\texttt{current}) > \tau_{d_i}$$

An important point to observe is that in this proposal the threshold for filter overriding is parametrized not just to the desirability function but to the particular determinant, as different determinants may have different thresholds.

Because it is sensitive to known determinants of desirability, this approach meets the computational constraint on reconsideration. Determining whether to reconsider does not entail paying the full cost of deliberation. Furthermore, the proposal for the filter override grounds reconsideration in responses that require minimal cognitive and computational effort on the part of the agent. For this reason, it coheres with Bratman's insight that it is best to model reconsideration as the product of non-deliberative tendencies. Needless to say, the proposal is different from Bratman's idea of relying on 'general habits and propensities'. Absent substantive assumptions, we cannot identify the agent's attitudes towards known determinants of desirability with general habits and propensities'. But even without a straightforward identification, the ideas behind these proposal might be fruitfully combined—perhaps by assuming that tracking known determinants is an example of such general habits and propensities.[17]

It should be relatively clear how the account is to be used in modeling reconsideration in the Tileworld task. We have noted that, in the simplest version of the task, point values of holes are known determinants; so, if a newly spawned hole has a point value that significantly exceeds the current goal, deliberation is to be reopened. This shows the Pollack and Ringuette account to be a special case of the present proposal: any example that can be modeled by a comparison of accumulated points can be modeled by the known determinants account. Another example which can be aptly modeled by the account is if the agent-tile-hole path for the newly spawned hole is significantly shorter than the one for the current goal. Hole distance may count as its own determinant, so holes that are particularly near might attract the agent's attention.

There may also be other, less obvious aspects of the agent's psychology that could be added on to the agent's description. These would enrich the variety

[17] One of the most interesting directions of expansion for these agent architectures is how they might be combined with neural networks to lead to hybrid systems of sorts. For instance, Thomason and Horty (2022) suggest in passing that the desire module in an IRMA-style architecture might be assigned to a neural network. In the same speculative mode as Thomason and Horty, one might supposed that there is potential to generate known determinants of desirability via a kind of neural network.

of ways in which the determinant comparison can be triggered in a deliberative context. Recall the variant of the task in which the agent's goal is to accumulate 100 points and no more than that. Imagine them sitting at 99 points and intending to fill a 3-point hole which is located 10 steps away. As they are moving, a new hole spawns that is worth 1 point and is 5 steps away. This new hole might not meet either the point-value or the proximity thresholds. However, if the agent's determinants are spelled out in a rich enough way, it will be possible to identify determinants that trigger reconsideration. For example, the agent's goal of accumulating 100 points can be turned into its own determinant of desirability, and so that too can be captured under the umbrella of the current proposal.

If a deliberative context includes unknown determinants of desirability—that is determinants that are not immediately known by the agent—they will not affect the filter override. This is as it should be: the stability-setting role of intentions should be strong enough that everyone but the most cautious agents should not be bogged down working out complex matters just for the purpose of filter overriding.

8 Formal presentation of the known determinant account

To highlight some choice points in spelling out the known-determinant account, I think it is important to give the account a formal sketch. I turn to this task in this concluding section.

Assume that deliberation happens against the background of a **deliberative state**. This is a triple $\delta = \langle \mathcal{B}, \mathcal{D}, \mathcal{I} \rangle$, where $\mathcal{B}$ is a credence function–a function from propositions forming a σ-algebra to real numbers;[18] $\mathcal{D}$ is a structured desirability function (more on this concept in the next paragraph); and $\mathcal{I}$ is a set of intentions, modeled as a set of options (also discussed in greater depth below). In principle, it would be good to keep separate tabs of those intentions that have been merely been committed to and those intentions that have been already executed. In practice, it simplifies our formalism to blur this distinction for the time being. As a point of notation, '$\mathcal{B}_\delta$', '$\mathcal{D}_\delta$' and '$\mathcal{I}_\delta$' are ways of denoting the individual coordinates of δ.

A **structured desirability function** $\mathcal{D}$ is a pair $\langle \langle d_1, ..., d_n \rangle, h \rangle$ consisting of a sequence $\langle d_1, ..., d_n \rangle$ of known determinants of desirability and an aggregation method h. (For simplicity, I assume that each determinant has the same type as the aggregated desirability function itself, mapping outcomes to real numbers.) How the aggregation method h combines the individual determinants is left untheorized

[18] We may require credence functions to be probabilistic, but this is irrelevant to the structural points I want to highlight.

for the present purposes. For one thing, in the indirect approach the known determinants are reverse-engineered *post-hoc*: so we should be careful to avoid reading h as meaning that these factors play a causal role in generating the relevant outcomes. For another, even in the direct approach, the matter could be as simple as there being a single determinant d_1 and h being the identity function: this would crown d_1 as the all-things considered desirability function. Alternatively, it might be that h performs some (possibly weighted) average of the known determinants in $\langle d_1, ..., d_n \rangle$; or it could be something else entirely. Indeed, since we did not include *unknown* determinants in $\langle d_1, ..., d_n \rangle$, h will have to incorporate any information about desirability that goes beyond what is known in the relevant sense. Thus in any example in which there are unknown determinants will require h to be much more complex than an average of known determinants.

We model options as pairs $\langle \Pi, \mathrm{K} \rangle$ consisting of a set of preconditions Π, and a set of consequences K.[19] Given an option, the preconditions and consequences of that option are themselves modeled as consistent sets of propositions. Write o's preconditions as 'Π_o' and o's consequences as 'K_o'. It is helpful to have a notion of incompatibility between options that is specified in terms of this preconditions/outcome structure. Multiple such notions are available. We could say that two options m and n are *outcome*-incompatible iff $\bigcap(K_m \cup K_n) = \emptyset$—that is, if their consequences are incompatible. To do the work that this is supposed to do, the consequences of any one option must be specified in a fine-grained way: suppose an agent has the option (o_1) of slicing the bread with their right hand and (o_2) of slicing the bread with their left hand. These options are intended to be incompatible, so K_{o_1} and K_{o_2} must include more than just the mere state that the bread is sliced. For another notion of incompatibility, we might say that m and n are *sequentially*-incompatible iff the consequences of m preclude the preconditions of n, i.e.$\bigcap(K_m \cup \Pi_n) = \emptyset$. Sequential incompatibility is important in the context of generating options at the beginning of deliberation, after an update. Here, however, in the interest of focus on the mechanics of reconsideration we will downplay the importance of this diachronic perspective.

We group together those options that are, in an intuitive sense, answers to the same question. Think of a question as a set of pairwise outcome-incompatible

[19] It is more standard Weld (1994); Horty and Pollack (2001) to model individual options in terms of the states of affairs they bring about—e.g. a propositions corresponding to the claim that a given event took place—and then connect these by 'causal links', which are propositions specifying the preconditions and consequences of the option. Here I find it simpler to identify options with preconditions and consequences. I am not sure whether this model is just a way of reshuffling the standard model around, or whether it is in some important sense non-equivalent.

options.[20] The function $\mathcal{Q}$ maps options to the question that they are answers to (so $\mathcal{Q}(o)$ is a set of outcome-incompatible options). The constraint that plans be consistent requires that for any question q and deliberative context δ, $\mathcal{I}_\delta$ records at most one answer to q. Designated this answer as $\mathcal{A}(q)$. If the plan records no answer to q, let $\mathcal{A}(q)$ map to some dummy object.

The **deliberation** function inputs a deliberative state δ and a set of options $\mathcal{Q}$, and outputs a state δ', which updates δ with some new committed intentions—i.e. an option from $\mathcal{Q}$ that was added to $\mathcal{I}_\delta$. In other words, $\mathrm{del}(\delta, \mathcal{Q}) \mapsto \delta'$ where δ' agrees with δ on $\mathcal{B}$ and $\mathcal{D}$ but may include an updated $\mathcal{I}$. Of course deliberation is not the only process that updates some elements of the agent's deliberative state. At each tick of the clock the agent may update their beliefs and desires.

At any step of deliberation, the set of relevant options is the union of two sets—the **live options** and the **override options**. The live options are those options that secure some live goal in δ. The **live goals** in δ are the preconditions of committed intentions in δ such that the agent does not assign credence 1 to their being satisfied, conditional on the execution of committed but not yet executed intentions. Note that the agent will perform appropriate credence updates every time they perform an action. In particular, the credence function is conditionalized on the consequences of actions that the agent has already executed. Formally, we write this as follows:

$$\textbf{live goals}_\delta = \{\varphi \mid \mathcal{B}_\delta(\varphi) < 1 \ \ \& \ \ \exists o \in \mathcal{I}, \varphi \in \Pi_o\}$$

To illustrate, suppose I have a committed intention to cook the pasta. The precondition of the intention is to have the pasta available in the first place, and that the agent does not already believe that this precondition is satisfied. Then having the pasta available will be a live goal. By contrast, if the agent were already certain that the target proposition was satisfied (i.e. $\mathcal{B}_\delta(\varphi) = 1$) then the goal cannot be live in the relevant sense.

Next, exploit the agent's live goals to define their **live options** (in δ) as follows:

$$\textbf{live options}_\delta = \{o \mid \exists \varphi \in \textbf{live goals}_\delta, \varphi \in K_o\}$$

Informally, the live options in context δ are those options that secure a live goal, i.e. some precondition of a committed intention. The difference between live goals

[20] For a defense of the idea that intentions might be sensitive to questions, see Beddor and Goldstein (ms.).

and live options is that the former are mere outcomes (e.g. having cooked pasta) while the latter are options available to the agent in the pursuit of their desired, and committed, outcomes (e.g. cooking pasta).

All of this is relatively transparent, if rather sketched out. Our central focus is to characterize the behavior of the filter override, which we do by characterizing a (possibly empty) set of **override options**. Recall that our vision is to identify those options which are not live, but are sufficiently better than some settled option with regards to (at least) one known determinant of desirability. Formally:

$$\textbf{override options}_\delta = \{o \notin \textbf{live options}_\delta \mid \exists n[d_n(o) - d_n(\mathcal{A}(\mathcal{Q}(o))] > \tau(d_n)\}$$

The component that requires the most in the way of explanation is the term $d_n(\mathcal{A}(\mathcal{Q}(o))$. Recall that $\mathcal{Q}(o)$ is the question that o answers. Recall also that given a question q, $\mathcal{A}(q)$ returns the option that is currently committed as the answer to q, if there is one. So $\mathcal{A}(\mathcal{Q}(o))$ returns the option that is currently committed to as the answer to whatever question o is an answer to. In other words: we are comparing o with some currently committed option that is an answer to the same question.

If the plan currently contains no option that is an answer to the same question as o, then I am suggesting that o ought to remain filtered out. After all, it seems plausible that fresh questions ought to be targets in deliberation as part of the ordinary process, as they become relevant, and not be introduced by the exceptional override process. It is for this reason, that I prefer the current, question-sensitive analysis of the override module to a more permissive analysis that would compare the alternate prospect o with any one committed intention in $\mathcal{I}$.

9 Sample Application

To illustrate the mechanics of the framework, I apply it to a simple Tileworld task. Suppose that at t_1 the agent a takes in the world depicted in Figure 2. They start with no committed intentions, so I only contains very broad constraints — for example, that the agent intends to maximize the total sum of accumulated points by filling holes. This is reflected in the desirability function, which in turn also features the multiple coordinates noted before. In particular, assume for simplicity that $\mathcal{D}$ has two dimensions: d_1 for hole value and d_2 for hole distance. As for the credence function $\mathcal{B}$, it only reflects two kinds of uncertainty: (1) uncertainty concerning when the two holes that are present on the board will disappear and (2) uncertainty concerning the possibility that some new holes might appear in the next turns.

In the initial round, there are only two holes available and consequently the agent can deliberate only between the options **fill2** (fill the 2-hole that is 4 tiles away) or **fill5** (fill the 5-hole that is 6 tiles away), so $O = \{\textbf{fill2}, \textbf{fill5}\}$. There is a value-distance tradeoff which different deliberation functions will resolve in different way, but let's suppose the present deliberation function resolves it in favor of **fill5**. In the next round, there is a committed intention, **fill5**, which excludes **fill2** from consideration and focuses the agent's attention to the preconditions of **fill5**. For the 5-hole to be filled, the agent must move toward it and fill each of its three tiles with a tile. These are the agent's current **live goals**, and so the agent's **live options** are those that secure these goals.

So far, so good. This is more or less the ordinary behavior of a *BDI* agent. Now for the distinctive contribution of this paper. Suppose that between the end of the previous round, and the beginning of this round, a new 7-point hole has sprouted in the south-western quadrant of the Tileworld. Should the agent reconsider their intentions? This will depend on the value of certain parameters. Consider then the option to fill the new hole (**fill7**). Like every other option, this can be assessed by the agent according to its structured desirability. (How many points is it worth? How valuable is it?). The relevant dimension here is point-value, and assume it denoted by d_1. We have $d_1(\textbf{fill7}) = 7$. This value is to be compared to $d_1(\mathcal{A}(\mathcal{Q}(\textbf{fill7}))$. Recall that the staggered operators $\mathcal{A}(\mathcal{Q}(\cdot))$ are meant to identify the currently settled answer to the question that is associated with the given option. The question that **fill7** answers is *which currently uncovered hole should I fill?*. Formally, we may think of this as the set of answers $\{\textbf{fill2}, \textbf{fill5}, \textbf{fill7}\}$. The current answer to this question is **fill5**. Thus, the relevant comparison is between $d_1(\textbf{fill7})$ and $d_1(\textbf{fill5})$. The difference here is 2 in favor of **fill7**. This will trigger reconsideration iff 2 is greater than the threshold $\tau(d_1)$—the threshold associated with point values.

10 Conclusion

The known determinant account I presented in the previous sections has the generality of the desirability comparison account, but clarifies the sense in which filter overriding is cheaper than deliberation, and yet manages to be an imperfect predictor of its outcomes. There remains further room to entertain different implementations of the account, and there is much urgency to compare it to some of the alternatives in the context of simulations, within the Tileworld paradigm and beyond.

As noted in passing, there is a systematic correspondence between proposals within a planning architecture and philosophical theories of instrumental rationality for "ordinary", biological agents. I explored one direction of this correspondence: I identified one debate in the philosophy of action and explored how it plays out when interpreted as a debate about the rational dynamics of an artificial agent. One key question that this development raises is what, if anything, this teaches us about the modeling of biological rational agents.

My case for the known determinant account rests on the failure of some alternatives to meet some key desiderata having to do with flexibility and accuracy of the reconsideration mechanism. The question whether a similar case can be made in the biological case depends on the underlying similarities between the cognitive structures of biological and artificial agents, and it is clearly too broad to be tackled in this concluding section.

However, I have also floated an idea that ought to carry over independently of cognitive similarities. One promising option I have considered for artificial rational agents is the idea that they might maintain at once a utility function and a quick-and-dirty model of their own utility function — inspired by the *post-hoc* explanatory algorithms in XAI. If this turns out to be the best way of thinking about reconsideration in an artificial planning agent, it seems exceedingly likely that it should be incorporated in our models for biological planning agents. One benefit of this move is that we would not have to rewrite standard static decision theory in terms of structured utility/desirability functions. Instead, we could limit our appeal to the agent's rough structured model of the desirability/utility to the context of specifying the details of the reconsideration module.

Bibliography

Babic, Boris, Gerke, Sara, Evgeniou, Theodoros and Cohen, I Glenn (2021), 'Beware explanations from ai in health care', *Science*, 373(6552), 284–286.

Beddor, Bob and Goldstein, Simon D (ms.), 'A question-sensitive theory of intention', National University of Singapore and Australian Catholic University.

Bratman, Michael E. (1987), *Intention, Plans, and Practical Reason*, Harvard University Press.

Bratman, Michael E. (1992), 'Planning and the stability of intention', *Minds and Machines*, 2(1), 1–16.

Bratman, Michael E. (2018), *Planning, Time, and Self-Governance*, Oxford University Press.

Bratman, Michael E., Israel, David J and Pollack, Martha E (1988), 'Plans and resource-bounded practical reasoning', *Computational Intelligence*, 4(3), 349–355.

Conlisk, John (1996), 'Why bounded rationality?', *Journal of Economic Literature*, 34(2), 669–700.

Easwaran, Kenny (2014), 'Decision theory without representation theorems', *Philosophers' Imprint*, 14, 1–30.

Evans, Jonathan St BT and Stanovich, Keith E (2013), 'Dual-process theories of higher cognition: Advancing the debate', *Perspectives on psychological science*, 8(3), 223–241.

Fleisher, Will (2022), 'Understanding, idealization, and explainable ai', *Episteme*, 19(4), 534–560.

Fletcher, Logan and Carruthers, Peter (2012), 'Metacognition and reasoning', *Philosophical Transactions: Biological Sciences*, 367(1594), 366–1378, metacognition: computation, neurobiology and function (19 May 2012).

Georgeff, Michael, Pell, Barney, Pollack, Martha, Tambe, Milind and Wooldridge, Michael (1998), 'The belief-desire-intention model of agency', in 'International workshop on agent theories, architectures, and languages', 1–10, Springer.

Ghallab, Malik, Nau, Dana and Traverso, Paolo (2004), *Automated Planning: theory and practice*, Elsevier.

Griffiths, Thomas L, Callaway, Frederick, Chang, Michael B, Grant, Erin, Krueger, Paul M and Lieder, Falk (2019), 'Doing more with less: meta-reasoning and meta-learning in humans and machines', *Current Opinion in Behavioral Sciences*, 29, 24–30.

Harman, Gilbert (1976), 'Practical reasoning', *The Review of Metaphysics*, 29(3), 431–463.

Harman, Gilbert (1986), *Change in View*, MIT Press.

Herzig, Andreas, Lorini, Emiliano, Perrussel, Laurent and Xiao, Zhanhao (2017), 'BDI logics for BDI architectures: old problems, new perspectives', *KI-Künstliche Intelligenz*, 31, 73–83.

van der Hoek, Wiebe, Jamroga, Wojciech and Wooldridge, Michael (2007), 'Towards a theory of intention revision', *Synthese*, 155(2), 265–290.

Holton, Richard (2009), *Willing, Wanting, Waiting*, Oxford University Press.

Horty, John F and Pollack, Martha E (2001), 'Evaluating new options in the context of existing plans', *Artificial Intelligence*, 127(2), 199–220.

Icard, Thomas (forthcoming), *Resource Rationality*, Cambridge University Press.

Icard, Thomas, Pacuit, Eric and Shoham, Yoav (2010), 'Joint revision of beliefs and intention', in 'Twelfth International Conference on the Principles of Knowledge Representation and Reasoning', .

Kinny, David and Georgeff, Michael (1991), 'Commitment and effectiveness of situated agents', in 'Proceedings of the Twelfth International Joint Conference on Artificial Intelligence (IJCA1- 91)', 82–88.

Krishnan, Maya (2020), 'Against interpretability: a critical examination of the interpretability problem in machine learning', *Philosophy & Technology*, 33(3), 487–502.

Lipton, Zachary C (2018), 'The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.', *Queue*, 16(3), 31–57.

Lorini, Emiliano and Herzig, Andreas (2008), 'A logic of intention and attempt', *Synthese*, 163(1), 45–77.

Millgram, Elijah (2019), 'Bounded agency', in Luca Ferrero (ed.), 'The Routledge Handbook of Philosophy of Agency', 68–76, Routledge.

Osman, Magda (2004), 'An evaluation of dual-process theories of reasoning', *Psychonomic Bulletin & Review*, 11(6), 988–1010.

Paul, Sarah (2020), *Philosophy of Action: A Contemporary Introduction*, Routledge.

Pollack, Martha E and Ringuette, Marc (1990), 'Introducing the tileworld: Experimentally evaluating agent architectures', in 'AAAI', volume 90, p183–189.

Ridge, Michael (1998), 'Humean intentions', *American Philosophical Quarterly*, 35(2), 157–178.

Russell, Stuart (2016), 'Rationality and intelligence: A brief update', *Fundamental issues of artificial intelligence*, 7–28.

Russell, Stuart and Norvig, Peter (1995), *Artificial intelligence: a modern approach*, Prentice hall, references are from the third edition (2010).

Russell, Stuart and Wefald, Eric (1991), 'Principles of metareasoning', *Artificial intelligence*, 49(1-3), 361–395.

Russell, Stuart J and Norvig, Peter (2020), *Artificial Intelligence a Modern Approach, 4th edition*, Pearson.

Schut, Martijn and Wooldridge, Michael (2000), 'Intention reconsideration in complex environments', in 'Proceedings of the fourth international conference on Autonomous agents', 209–216.

Schut, Martijn and Wooldridge, Michael (2001), 'Principles of intention reconsideration', in 'Proceedings of the fifth international conference on Autonomous agents', 340–347.

Schut, Martijn, Wooldridge, Michael and Parsons, Simon (2004), 'The theory and practice of intention reconsideration', *Journal of Experimental & Theoretical Artificial Intelligence*, 16(4), 261–293.

Sloman, Steven A (1996), 'The empirical case for two systems of reasoning.', *Psychological bulletin*, 119(1), 3.

Tenenbaum, Sergio (2018), 'Reconsidering Intentions', *Noûs*, 52(2), 443–472.

Tenenbaum, Sergio (2020), *Rational Powers in Action: Instrumental Rationality and Extended Agency*, Oxford University Press.

Thomason, Richmond H and Horty, John (2022), 'Artificial and machine agency', in Luca Ferrero (ed.), 'The Routledge handbook of philosophy of agency', 366–375, Routledge.

Van Zee, Marc and Icard, Thomas (2015), 'Intention reconsideration as metareasoning', University of Luxembourg and Stanford University; available at https://web.stanford.edu/ icard/borm.pdf.

Velleman, Daniel J. (1989), *Practical Reflection*, Princeton University Press.

Weld, Daniel S (1994), 'An introduction to least commitment planning', *AI magazine*, 15(4), 27–27.

Wooldridge, Michael (2009), *An introduction to multiagent systems*, John wiley & sons, 2nd edition (1st edition, 2002).