Completing Bratman's Intention

January 17, 2025

Abstract

In his What is intention, Bratman proposed that intentions can be seen as both inputs and outputs of practical reasoning. However, he chose not to pursue a full theory to distinguish which outputs of practical reasoning are intentions and which outputs are not. Building on Bratman's analysis of intention, we develop a sequence of theories. An initial naive theory suggests that an event is intended if it cannot be excluded without undermining the agent's goal. However, this approach fails in non-trivial cases with causal dependencies. A revised theory is then offered to incorporates an agent's subjective understanding of causal relationships to better align intentions with practical reasoning. Finally, the theory is further extended to account for probabilistic beliefs, defining intentions as commitments to altering event probabilities based on an agent's reasoning process. The resulting framework is evaluated through thought experiments, addressing limitations in earlier approaches and capturing the stability and practical role of intentions. The paper concludes by situating the theory as a step towards completing Bratman's work on intention.

1 Introduction: the task Bratman left to us

In his *What is intention*, Bratman proposed that intentions can be seen as both inputs and outputs of practical reasoning. (Bratman, 1990, p.29) He also considered a thought experiment, which he calls "the problem of the package deal", from which he concluded that not all of the outputs of practical reasoning are intended by an agent. (Bratman, 1990, p.30) However, he chose not to pursue a full theory to distinguish which outputs of practical reasoning are intentions and which outputs are not. This paper attempts to give such a theory founded on Bratman's take on intentions. Before presenting it, we shall begin by briefly introducing the package deal problem and Bratman's take on it in Section 2. The main theory will be developed in Section 3 and generalized to accommodate probabilistic beliefs in Section 4. Section 5 concludes the paper by briefly arguing that the theory captures the characteristics of intention desired by Bratman.

2 Package Deal Problem & Choice-Intention Principle

Bratman introduced the problem of the package deal via a thought experiment concerning bombers (Bratman, 1990, p.23): A Terror Bomber (hereafter TB) and a Strategic Bomber (hereafter SB) both intend to pursue their goal of winning the war by dropping bombs. TB plans to bomb the school in enemy territory, killing their children, terrorizing their population, forcing the enemy to surrender, and thereby win the war. On the other hand, SB plans to bomb the enemy's munitions plant, undermining their war effort, and achieve victory. In addition, SB clearly knows that by bombing the munitions plant he will destroy the school next to the plant and therefore kill all the children inside the school. However, he decides that the cost of the children's life is outweighed by the potential good outcome of destroying the munition plant, so he will bomb it anyway.

Bratman wanted to say that TB intends to kill the children, while SB does not. However, in choosing to bomb the munition plant, SB knows that he will kill the children. So, SB choosing to bomb is like opting for a package which includes both destroying the munition plant and killing the children. How, then, can SB be said to not intend killing the children, if this is a part of his deliberate choice? Bratman's goal is to provide an argument against the apparently unavoidable conclusion that SB does so intend.

Bratman articulates the premises of the package deal argument, one of which is the choice-intention principle: "If on the basis of practical reasoning I choose to A and B and to . . . (A and B are outputs of practical reasoning), then I intend to A and B and to ... " (Bratman, 1990, p.25) He then rejects this principle, based on his analysis that intentions play an important role in future reasoning and action (intentions must also be inputs of practical reasoning). For TB, his intention to kill the children is a conduct controller that guides him to kill them: If midair he finds out the children have moved to a different place, he will try to track them down and bomb the new place instead. If a troop movement will result in evacuation of the children, then TB would not allow the troop

movement. On the other hand, SB will not be bothered midair if he finds out that the children have moved, and he will probably be happy with the fact that he doesn't need to kill the children. He will also not worry about the fact that troop movements might cause the children to evacuate. Therefore, SB does not intend to kill the children, because his choice of killing the children does not play the roles that intentions should play. (Bratman, 1990, p.27)

According to Bratman, choices are made after holistic considerations taking factual constraints and rational reasoning into account, while intentions do not need to be holistic. That explains why SB can intend to only bomb the munition plant but choose to bomb the plant and kill the children. However, we cannot deny that intentions are also influenced by factual constraints and rational reasoning. For example, SB's intention to bomb the munition plant is also dependent on the fact that munition plants are important to the enemy's war efforts, and SB's reasoning that destroying it will have an negative impact on the enemy. Bratman admits, in the end of his paper, that more work needs to be done (Bratman, 1990, p.30):

> Of course, having rejected the choice-intention principle, we will need to put something in its place. One's favor of an overall scenario will involve one's coming to have some intentions or others-intentions that will guide future reasoning and action... A full theory will need to say more about *which* intentions an agent will thereby come to have; but I cannot pursue this matter further here. Suffice it to say that an acceptable treatment of the relation between intention, choice, and practical reasoning cannot just assume that one intends to all of what one chooses on the basis of practical reasoning.

The remaining parts of this paper will attempt such a treatment, taking inspirations from the package deal problem.

3 Developing the Theory

Let us consider again why intuitively we think that SB intends to bomb the plant, while he[1] does not intend to kill the children. Despite him deciding to bring about the death of the children, he would not let that happen if hypothetically he had a way to destroy the plant and leave the school safe. On the other hand, he certainly would choose to bomb the plant if he were given the option to not bomb it. After all, the first choice he made was to bomb the plant. So, we can make a naive first attempt at saying which of the choices an agent makes are intentions:

> **Theory – Naive Attempt**
>
> Given a goal G and a set[2] E (package) of events which an agent A decides (through practical reasoning) to bring about in order to achieve G, an event $e \in E$ is intended by A towards G if and only if A would hypothetically choose to bring about E over bringing about only $E \setminus \{e\}$.

We can then say that an event e is an intention of an agent A if and only if there exists a goal G such that e is intended by A towards G.

Looking at the definition, some may argue that what people intend should be actions rather than events. This should not be a big concern, because actions can be translated into events. For example, the sentence "I intend to have a cup of tea" can be translated into "I intend to bring about the event 'I have a cup of tea'." Although somewhat unnatural, this translation does not change the meaning of the natural sentence. In this paper, we take events to be what an agent can intend, so that we have a more expressive language for intentions in complex scenarios like the bombers' story, in which an agent could intend to bring about events like "the enemy surrenders". Another

[1] We adopt the same pronouns that Bratman uses for the bombers.

[2] The naive concept of set and standard set-theoretic notations are used throughout this paper. Explanations of the concept of naive set and meanings of notations in naive set theory can be found in almost all logic textbooks, for example, Appendix A of *Mathematical Logic and Computation* (Avigad, 2022).

feature of the theory is that the existence of an intention requires the existence of a goa. But is this always the case? It is not hard to argue that we sometimes have spontaneous intentions that might not appear to serve a specific goal. In cases like this, we can be flexible and say that the goal is to "temporarily follow my heart" or "do whatever I want for now".

Informally, the naive attempt says that an event in a package which an agent chooses to bring about is an intention if they would choose the original package over the package with the event removed. Let us test whether our theory identifies simple everyday intentions. Suppose that I feel thirsty and decide to have a cup of tea in order to make myself not thirsty. Does our theory capture my intention of having a cup of tea? Here the goal G is to make myself not thirsty, and the package E I choose to bring about is the singleton set $\{e\}$, in which e is the event that I have a cup of tea. The theory suggests that I hypothetically choose between bringing about E and bringing about $E \setminus \{e\} = \emptyset$, and if I prefer the original package, then I do intend e. Clearly, I would choose to have a cup of tea over doing nothing, so the theory successfully identifies my intention to have a cup of tea.

However, we can see an immediate flaw of our theory applied in the bomber story: Say for TB the set (package) of events he decides to bring about is $\{b,k,t,s\}$ towards the goal of victory, in which b is the event that TB drops the bomb on the school, k is the event that the children in the school are killed, and t is the event that the enemy people are terrorized, and s is the event that the enemy surrenders. Intuitively, all of the four events should be TB's intentions. Let us apply our naive theory to test whether k is an intention. Comparing just $\{b,t,s\}$ happening with $\{b,k,t,s\}$ happening, it does not seem like TB would prefer the original package (unless he enjoys watching children die), because s happening is enough for him to win. A similar analysis shows that TB's intentions b and t are also not captured by our naive theory.

So, what is wrong? Note that our theory successfully identifies s as TB's intention: He will clearly prefer the original package $\{b,k,t,s\}$ over just $\{b,k,t\}$ towards achieving his goal, because he

relies on s happening to win the war. What our theory fails to capture in the case of k is the fact that TB relies on k happening to cause t, and then s. In removing k from the package and leaving its consequences t and s inside, we broke the causal chain that led TB to have the intention k, leaving the chain disconnected. So, it might help to make sure that causal chains remain connected, by adding the constraint that when removing an event from a package, all of the events that follow from it causally must also be removed. Let $C(e,E)$ denote the set of events in E that follow causally from the event e, plus e itself. We can revise our naive theory to the following:

> **Theory – Second Attempt**
>
> Given a goal G and a set E of events which an agent A decides to bring about in order to achieve G, an event $e \in E$ is intended by A towards G if and only if A would hypothetically choose to bring about E over only bringing about $E \setminus C(e,E)$.

We still say that an event e is an intention of an agent A if and only if there exists a goal G such that e is intended by A towards G, and this will remain unchanged.

We can verify that our new theory correctly identifies TB's intentions: s is still an intention because $C(s,\{b,k,t,s\}) = \{s\}$, so the new theory agrees with the naive theory on s. Consider k from $\{b,k,t,s\}$ again. Because t and s causally follows from k while b does not, we have $C(k,\{b,k,t,s\}) = \{k,t,s\}$, and then $\{b,k,t,s\} \setminus \{k,t,s\} = \{b\}$. Between $\{b\}$ and $\{b,k,t,s\}$, TB would clearly prefer $\{b,k,t,s\}$, the original package which allows him to win. Therefore, our new theory correctly identifies k as TB's intention. The analysis for b and t goes through similarly.

The TB scenario is simple in the sense that all the choices that he made are his intentions. Can our theory perform equally well on SB's choices, in which one of them is not an intention? Say that SB decides to bring about the package $\{b,d,k,s\}$ towards his goal of victory, in which b is the event that SB drops the bomb on the munition plant, d stands for the destruction of the munition plant, and k, s the same as in TB's scenario. We want to see whether our theory successfully says

that k is not SB's intention. Apparently, k does not cause b or d. But here we see something subtle, because k could terrorize the enemy population as it did in the TB case, and therefore cause s. Taking this into account and applying our new theory, $C(k,\{b,d,k,s\}) = \{k,s\}$, and SB certainly would choose the whole package $\{b,d,k,s\}$ over $\{b,d,k,s\} \setminus \{k,s\} = \{b,d\}$, because SB needs the enemy to surrender. Therefore, our theory incorrectly suggests that k is SB's intention.

One problem here is that SB is not aware of the causal relationship between k and s, while the theory concerns objective causal relationships. Adopting Bratman's analysis that intentions are outputs of the agent's practical reasoning, it is clear that the causal relationships between events should be analyzed from the agent's perspective. So, we should replace $C(e,E)$ with $C^A(e,E)$ in our second theory, where $C^A(e,E)$ is the set of events in E that the agent A considers to causally follow from the event e, plus e itself. With this replacement, $C^{SB}(k,\{b,d,k,s\}) = \{k\}$, and SB, not wanting to kill the children, will pick $\{b,d,k,s\} \setminus \{k\} = \{b,d,s\}$ over the whole package $\{b,d,k,s\}$. Therefore, k is not intended by SB towards his goal of victory. Supposing that SB does not have other goals that will lead him to decide having the children killed, we can say that SB does not intend k. One can say that s, although being an intention towards victory, is also one of SB's goals, and he decides to bring about $\{b,d,k\}$ in order to achieve s. But in this case, TB will choose $\{b,d\}$ over $\{b,d,k\}$, so k is not intended by SB towards d, and therefore our conclusion that k is not SB's intention still holds. We can also apply our theory to confirm that b, d, s are intended by SB. It seems to work well after the fix.

However, a slightly tweaked version of the above scenario poses another threat: Suppose that SB still has the goal of victory, for which he plans to bomb the munitions plant (b), destroy it (d), and force the enemy to surrender (s), thereby leading to victory. However, he took some extra considerations to realize that bombing the munitions plant will kill the children nearby (k), which will terrorize the enemy population (t), and could also force the enemy to surrender (s). he does not want to achieve victory in this way. However, he decides again that the cost of the children's

lives is outweighed by the benefit of winning the war, so he decides to bring about the package $E = \{b, d, k, t, s\}$ in order to achieve victory. Similar to the previous SB scenario, he should not be considered to have the intention k. Yet this time SB knows that t and s causally follow from k, so $E \setminus C^{SB}(k, E) = E \setminus \{k, t, s\} = \{b, d\}$ which does not contain s, and SB would prefer the original package, suggesting that the revised version of our second theory fails here.

Why did our theory fail? Notice that in SB's plan, although s causally follows from k, it might still follow from d without k. In applying our theory to construct the comparison package to test whether k is intended by SB, maybe we should only remove, from SB's perspective, the causal influences of k on E, rather than removing all events in E that SB would see as causally follow from k. In that way, k and t will be removed and s will remain in the comparison package, allowing SB to prefer the comparison package over the original package. Let $C^{A}_{remove}(e, E)$ denote the set of events obtained by removing, according to an agent A's beliefs, an event e and the causal influences of e from a set of events E. Here is the updated theory:

> **Theory – Third Attempt**
>
> Given a goal G and a set E of events which an agent A decides to bring about in order to achieve G, an event $e \in E$ is intended by A towards G if and only if A would hypothetically choose to bring about E over only bringing about $C^{A}_{remove}(e, E)$.

We can test whether this third theory correctly excludes k from SB's intentions in the tweaked SB scenario. Recall that $E = \{b, d, k, t, s\}$. We remove, from SB's perspective, the causal influence of k on E, assuming that SB has the same common causal sense as us: b and d are clearly not causally influenced by k, because they precede k. t would not happen if k does not occur, so t should be removed. Although k has an influence on s, it might still happen even if k does not happen, because d might lead to it. Therefore, we end up with $C^{SB}_{remove}(k, E) = \{b, d, s\}$, and SB would prefer this package over the original one since it suffices to have s for the victory. Our theory successfully concludes that k is not intended by SB towards victory, and as long as k is not intended by SB

towards other goals, our theory says that SB does not intend to kill the children. It is not difficult to verify that this theory works well in all previous scenarios. In all those simpler scenarios, the causal influences of an event e are just the events that follow causally from it, from the agent's perspective. So, the third theory should agree with the revised second theory, which worked well on all of them. Have we found a perfect theory of intentions?

Consider SB's situation once again. Although he does not want to achieve his goal by killing the children, he can still believe that killing them increases the chance that the enemy will surrender. Removing the causal influence of killing the children should therefore remove that effect as well, but our theory is not capable of doing that. We will explain and address this problem in the next section.

4 Managing Probabilistic Beliefs

So far, we have been working under an ideal assumption that agents have certain beliefs about which events they can make happen, and which events causally follow from others. In reality, the bombers are likely aware that there is a chance their operations might fail. Similarly, it is probably the case that neither SB nor TB believe for sure that killing the children will lead to the enemy's surrender. So, a general theory of intentions cannot talk just in terms of an agent's belief of one event causally following another, but must manage the probabilistic aspect of an agent's belief of both outcomes and causal relations.

This section will not deal with objective probabilities of events or objective causal relationships between events because they are not what matters when it comes to intentions. As outputs of practical reasoning, whether something is an intention of an agent should depend on their genuine subjective beliefs of probabilities and causal relationships. In fact, we do not even have to commit to the objective existence of causality, we only need agents to believe in them. This section simply tries to offer a method for distinguishing intentions from packages when an agent has made decisions based on their subjective probabilistic beliefs about events and causal relationships. Critics may argue that basing identification of intentions on agent's subjective beliefs makes it difficult for applications in areas such as legal practice. But let me point out that in practice, we can always take a fixed "common sense" approach to probability and causalities instead, rather than trying to find out people's genuine subjective beliefs. In this section, we will assume agents have a classical, naive view of probabilities of events, but try to keep the theories developed generalizable.

To begin with, let us take the perspective of TB. Although TB may not be 100% sure that his operation is going to be successful, he still wants the events b, k, t, s to happen in choosing his package, having a subjective belief that these are likely to happen if he attempts the operation. So we can say that the output of his practical reasoning is actually the events that he decides to make

happen, paired with his estimate of the probability that the event is happening. For convenience, we call such a pair a probabilistic event:

> **Definition – Probabilistic Event**
>
> A probabilistic event is a pair (e, p) where e is an event and $0 \leq p \leq 1$.

The probability in a probabilistic event is always a probability from the perspective of an agent. Saying TB decides to bring about the probabilistic event $(b, 80\%)$ is equivalent to saying TB believes that his operation will have a 80% chance of successfully dropping the bomb on the school and decides to go for it.

Removing the causal influence of a probabilistic event from a package would alter the probabilities of other probabilistic events. For example, suppose that TB has the following beliefs:

- His operation has a 80% chance of successfully dropping the bomb on the school. Otherwise, the school will not be bombed (no other bombers are bombing the school).
- If the school is bombed, there is a 50% chance that children will be killed. Otherwise, the children will not be killed.
- If the children are killed, there is a 50% chance that the enemy's people will be terrorized. Otherwise, the people will not be terrorized.
- If the enemy's people are terrorized, there is a 50% chance that the enemy will surrender. Otherwise, the enemy will not surrender.

And with a naive calculation of probabilities based on his beliefs, TB decides to carry out his operation in pursuit of the probabilistic package $E = \{(b, 80\%), (k, 40\%), (t, 20\%), (s, 10\%)\}$, hoping for victory. Suppose also that in his perspective, removing the causal influence of $(k, 40\%)$ would amount to completely removing $(t, 20\%)$ and then $(s, 10\%)$, resulting in $C^{TB}_{remove}((k, 40\%), E) = \{(b, 80\%)\}$. We can update the last theory in Section 3 using probabilistic events:

Theory – First Probabilistic Attempt

Given a goal G and a set E of probabilistic events which an agent A decides to bring about in order to achieve G, a probabilistic event $(e,p) \in E$ is intended by A towards G if and only if A would hypothetically choose to bring about E over only bringing about $C^A_{remove}((e,p),E)$.

This theory is a preserving generalization of the old one, in the sense that we can recover the old theory by setting p to 1 in all probabilistic events. Therefore, both theories should give the same answers to all previous thought experiments upon rephrasing. Intentions can still reside in events rather than probabilistic events: Even if TB believes that there is only a 20% chance of successfully terrorizing the population, it can still be his intention to do so. We can say that an event e is an intention of an agent A if and only if there exists a goal G and a probabilistic event (e,p) whose first coordinate is e, and e is intended by A towards G.

Consider the theory with the TB example above. If he wants to achieve victory through killing the children and terrorizing the enemy population, he would prefer the package $\{(b,80\%),(k,40\%),(t,20\%),(s,10\%)\}$ over just $\{(b,80\%)\}$, so k is TB's intention according to our theory. On the other hand, suppose that based on his beliefs, SB decides to pursue the probabilistic package $F = \{(b,80\%),(d,80\%),(k,40\%),(t,20\%),(s,30\%)\}$. Moreover, assume that according to SB's beliefs, $C^{SB}_{remove}((k,40\%),F) = \{(b,80\%),(d,80\%),(s,20\%)\}$. If SB does not want to achieve victory through killing the children, he would prefer the package $\{(b,80\%),(d,80\%),(s,20\%)\}$ over $\{(b,80\%),(d,80\%),(k,40\%),(t,20\%),(s,30\%)\}$, accepting the drop in the chance of victory. Our theory then says that k is not SB's intention.

Notice, however, that we had to make assumptions such as "the event will not happen otherwise" in the TB example. Indeed, it is somewhat unclear what it means for an agent to decide to "bring about" a probabilistic event when that event could still occur by chance, even without any action

from the agent. Suppose TB thinks that there is a 10% chance that the enemy will surrender without the operations of TB, and by bombing the school and terrorizing the enemy people he can increase it to 20%. When TB commits to his plan, it is better to say that he is committed to bring about a change in the probability that s happens, rather than saying that he is committed to bring about s. We use the following definition to describe changes in the probabilities of events (the chance of s happening increases by 0.1[1]) based on a default probabilistic event (the chance of s happening is 0.1).

> **Definition – Event Probability Change**
>
> Given a default probabilistic event (e, p), an event probability change with respect to it is defined as a pair (e, c) where $0 \leq c + p \leq 1$.

Given a package of event probability changes with respect to some default probabilistic events, causal influences of an event probability change can be removed in a similar way as before. Suppose that SB believes the following:

1. All events $\{b, d, k, t, s\}$ have a 0.01 probability of happening by default[2].
2. By carrying out his operation, the chance that b, d, k, t, s happens will increase to 0.8, 0.8, 0.4, 0.2, 0.3 respectively. That is, he will bring about the event probability changes $(b, 0.79), (d, 0.79), (k, 0.39), (t, 0.19), (s, 0.29)$.
3. The increase in the chance of k causally contributes to all the increase in the chance of t, and together contributes 0.1 to the increase in the chance of s.

By deciding to bomb the munitions plant in order to achieve victory and recognizing that children

[1] From here, we use decimals instead of percentages for probabilities, to avoid ambiguities between absolute change and relative change.

[2] Here we can take that SB consider the default probabilities as the probabilities of these events occurring if SB decides not to attempt the bombing. In general, agents might have different default actions in mind. In such cases, an agent might take the default probabilities to be the probabilities of relevant events occurring if those default actions are carried out by the agent.

might get killed, SB is committed to the event probability change package $E_c = \{(b, 0.79), (d, 0.79), (k, 0.39), (t, 0.19), (s, 0.29)\}$. Based on the third belief, $C^{SB}_{remove}((k, 0.39), E_c) = \{(b, 0.79), (d, 0.79), (s, 0.19)\}$. We now present our final theory:

> **Theory – Final Attempt**
>
> Given a goal G and a set E_p of probabilistic events which an agent A believes to be relevant events attached to their default probabilities. Suppose A decides to bring about a set E_c of event probability changes with respect to elements of E_p in order to achieve G. Then, an event probability change $(e, c) \in E_c$ is intended by A towards G if and only if A would hypothetically choose to bring about E_c over only bringing about $C^{A}_{remove}((e, c), E_c)$.

When (e, c) is intended by A towards some G and c is positive, we can say that A intends for e to happen, or simply e is A's intention. For negative c we can say that A intends to stop e from happening. If SB prefers $\{(b, 0.79), (d, 0.79), (s, 0.19)\}$ over the original change package in the example above, our theory says that k is not an intention of *SB*, as we would expect. Our final theory is a preserving generalization of the previous attempt, because we can recover it by requiring all the probabilities in E_p to be zero. Then, a event probability change in E_c can be treated as a probabilistic event. Therefore, the final theory works just as well in all the previous examples.

Some might point out that, in practice, agents do not think of probability and causal influence in terms of numbers. They often only have vague judgments like "this is very unlikely to happen" and "there is a good chance that this will cause that". Since our theory only considers probabilities and causalities from an agent's perspective, it can easily adapt to an agent's subjective concepts of probabilities and causal influences. For example, the probability values in the theory (and supporting definitions) can be replaced with judgments such as "unlikely to happen", "absolutely happening", etc. Change can then be characterized in terms of the difference between these judgments, such as "decrease from very likely to happen to unlikely to happen".

5 Conclusion: what Bratman wanted

Taking Bratman's task, we have developed a fairly general theory to identify intentions of agents, offering a solution to the Package Deal Problem and a replacement for the Choice-Intention Principle. There could be many more ways to test and improve our final theory, but I believe it is on the right track of capturing the idea behind Bratman's intentions.

Why would Bratman like our theory? In the concluding remarks of *What is Intention*, he said that in addition to being outputs of practical reasoning, "Intentions are relatively stable attitudes that function as inputs to further practical reasoning" (Bratman, 1990, p.30). Our theory requires intentions to be outputs of practical reasoning: Intentions must be events which an agent decides to bring about in order to achieve a goal. Our theory allows intentions to be inputs to practical reasoning: An existing intention often becomes one of the agent's goals and generates more intentions. Consider TB, he could develop the intention to terrorize the enemy population first, and towards this goal, develop the intention to kill the children. Most importantly, our theory guarantees that intentions are stable: We require that the agent still prefer the original package, including the intention and its causal effects, even when given the chance to get rid of them while preserving other parts of the package. Together, I believe our final theory is close to the one that Bratman wanted.

References

Avigad, J. (2022). *Mathematical logic and computation*. Cambridge University Press.

Bratman, M. E. (1990, 06). What Is Intention? In *Intentions in Communication*. The MIT Press. Retrieved from https://doi.org/10.7551/mitpress/3839.003.0004 doi: 10.7551/mitpress/3839.003.0004