North Carolina State University Department of Computer Science
Technical Report TR-2010-20

Toward a quantitative theory of belief change: Structure, difficulty, and likelihood

(A progress report)

Jon Doyle[*] JON_DOYLE@NCSU.EDU
Department of Computer Science, North Carolina State University
Raleigh, NC 27695-8206 USA

Abstract

Effective persuasion requires identification and selection of points of leverage in the belief system of the person or population being addressed. This paper describes a model in which personal beliefs are held and changed through a combination of reasoned, rational, and reflexive means, and in which the same attempts at persuasion can yield different outcomes in different people. The model is used first to interpret a logic of possible changes of belief, and second to develop mechanical and economic measures of the difficulty and probability of different changes of belief.

∗. This research was supported in part by AFOSR MURI Grant No. A9550-05-1-0321 through a subcontract to the Massachusetts Institute of Technology for work on "Formal Models of Belief Change in Cultural Context". The author thanks Whitman Richards, Rajesh Kasturirangan, Andrew Wicker, and the MURI meeting participants for helpful discussions on the topic.

 Version of September 7, 2010 at 11 : 20

Contents

1. The problem of changing minds

Sometimes it is easy to persuade someone to change his mind; sometimes it is hard. Contrast "Brave, brave, Sir Robin" with "Hier stehe ich nicht; ich kann nicht anders."

If one has no idea why someone thinks the way he does, one can only use appearances or other knowledge to identify populations to which the person belongs and then try methods based on models of how people think who belong to those populations. The more one knows about the specifics of how a person thinks, the more leverage one can bring to bear in choosing methods that will best influence that individual. However, persuading someone to change how he thinks can be very difficult despite detailed knowledge of why he thinks the way he does. Indeed, changing one's own beliefs can be very difficult, even for people seemingly very aware of why they believe as they do.

In line with these observations, we develop a mathematical model of belief and belief change that provides formal means for comparing and measuring the difficulty of different changes and the likelihood of different changes via different means for influencing beliefs. Applying these measurements to models of the beliefs specific to particular populations or individuals of known or assumed character, one can identify changes of minimal expected effort, to gauge risk by identifying changes of minimal and maximal effort, to distinguish changable beliefs from unchangeable ones, and to identify the best point of leverage for producing a desired change. We do not address here the problem of how one identifies the belief or character of a specific individual or population, a problem already studied in decision analysis and social psychology with methods adaptable to the model presented here.

1.1 Approach

Our model involves a rich conception of belief and belief change in which belief states can exhibit some ambiguity about what is actually believed, and in which beliefs can be influenced by rational incentives for change, by new information that changes assumptions or triggers predictable habits of thought, and by motivations internal to the individual. To address these aspects, our model involves desires, preferences, intentions, motives, habits, and logical, structural, and habitual connections in characterizing possible belief states and changes. The structures considered here encompass supports or foundations of beliefs, the inherent constitution of mental and belief states, and the organization of belief states into locales or substates corresponding to mental subagencies. The preferences considered here differentiate beliefs and changes according to their sources, content, foundations, and history, among other things.

The main elements of the model developed here are quite abstract and cover a considerable range of models of belief. We augment these with some more specific elements that remain quite abstract but support several important means for analyzing the difficulty and likelihood of belief changes, and that point the way for how to extend the model to handle other models of belief.

We concentrate on formalizing those aspects of the mental substructures that generate entrenchment and resistance to change. These aspects include models of reasoned inferential policies that capture common patterns for justifying existing beliefs, as well as preferential and game-theoretic policies that capture common decisions reflecting and reinforcing existing beliefs. These models take into account second- and higher-order preferences and goals of the individual regarding its beliefs, actions, and relations to different social groups. We do not assume that all individuals think

in the same way, and our model provides means for characterizing individuals who differ in, for example, constitutive degrees of mental integrity and inferential capacity.

1.2 Background

The model presented here draws on numerous extant theories. Epistemology and economic theories study belief change in terms of idealized conceptions of logical or economic rationality that presume epistemic omniscience and consistency, as in modal logics of belief and knowledge [33]; instantaneous and complete assimilation of new information, as in Bayesian conditionalization [52]; and effortless, instantaneous deliberative reasoning in microeconomic and game-theoretic theories of individual and group action [36, 57]. Despite their ideal character, however, these theories have limitations. Bayesian updating, for example, does not address learning of new information that contradicts current beliefs. That case is addressed by theories of epistemic entrenchment [31], in which one belief is more entrenched than another if it is given up before the other when faced with contradictory information, but these theories in turn do not seek to characterize why or how one belief becomes more entrenched than another, and so provide little or no predictive power.

The investigation reported here attempts to study a more realistic conception of change in which the degree of logical and economic rationality is limited by structural or informational properties of the believer. The aim is to better understand belief and belief change when inconsistency and ignorance can persist with indefinite duration, in which learning takes time, and slows as habits accumulate, in which reasoning requires concentrated effort, in which change requires motivation and acceptance, and in which resistance to change can be passive or active. This portion of the model draws on prior work on truth or reason maintenance systems (TMS/RMS) [9, 20]; on related theories of nonmonotonic logics [41] and reasoned assumptions [11, 19]; on theories of economically rational belief revision [17, 16]; on theories of reflective and argument-oriented dialectical deliberation [7, 10, 13]; and on cognitive mechanics [21, 22]. We also draw on work in progress by Wicker [60] to formalize the notion of influence mechanism, by which individuals or groups engender different kinds of changes on the beliefs of others.

2. Believers

2.1 Individuals

We begin by considering a set $\mathcal{U}$ of *individuals*. For the moment, we make no assumptions regarding the nature of these individuals, whether they be humans, legal persons, or artificial agents. The following will focus mainly on finite sets of individuals, but many aspects of our formalization also cover the case of infinite sets of individuals, as are studied in economic theories of markets with a continuum of individuals [3].

We assume that each individual falls into one or more of a finite set of *types* $T_1, \ldots, T_n$. Within $\mathcal{U}$, we identify the individuals of type T by $T(\mathcal{U})$, which we also denote $\mathcal{U}_T$. We write $T \sqsubseteq T'$ when T is a subtype of T'. We require that $T(\mathcal{U}) \subseteq T'(\mathcal{U})$ when $T \sqsubseteq T'$, and that $\sqsubseteq$ is transitive.

2.2 Groups

We regard a *group* $g \subseteq \mathcal{U}$ as a set of individuals. The set $\mathcal{U}$ itself constitutes the largest group, and the empty set of individuals $\emptyset$ constitutes the smallest group. We sometimes treat singleton groups $g = \{i\}$ as if they were the individuals they contain.

We write $\mathcal{G}$ to denote the set of all groups. More generally, we follow Noll's notion of material universe [43] or materially-ordered set [45] and assume that $\mathcal{G}$ forms a Boolean lattice of subsets of $\mathcal{U}$ in which the lattice meet and join of groups g and g' are respectively the maximal group contained in and the minimal group containing both of them. We call the lattice complement g^{e} the *exterior* or *social environment* of g.

For finite $\mathcal{U}$, we normally assume that every subset of $\mathcal{U}$ forms a group, so $\mathcal{G}$ consists of the power set $2^{\mathcal{U}}$. In this case, meet and join are just set intersection and union, respectively, and $g^{\mathrm{e}} = \mathcal{U} \setminus g$.

For infinite $\mathcal{U}$, we often will assume that $\mathcal{U}$ carries a topology and concomitant notions of topological interiors and closures. Different assumptions about the nature of groups motivate different topological definitions of groups. If singleton sets of individuals are to be degenerate groups, it is natural to regard $\mathcal{G}$ as the set of all subsets of $\mathcal{U}$, that is, to assume that $\mathcal{U}$ carries the discrete topology. If groups are conceived as containing all the neighbors of each individual in them, one might regard $\mathcal{G}$ as the set of regularly open subsets of $\mathcal{U}$, that is, sets which are the interiors of their closures, that is, $g = \operatorname{int}\operatorname{clo} g$.

As with individuals, we assume that each group falls into one or more group types $T_1, \ldots, T_m$, and write $T \sqsubseteq T'$ to mean that T is a subtype of T'. We write $T(\mathcal{G})$ or $\mathcal{G}_T$ to denote the set of groups of type T, and require that $T(\mathcal{G}) \subseteq T'(\mathcal{G})$ if $T \sqsubseteq T'$.

2.3 Group individuals

It is sometimes natural to regard the individuals of some type as forming a group. Conversely, it is also natural to regard some individuals, such as legal persons, corporations, organizations, and clans, as consisting of groups of other individuals. We thus partition $\mathcal{U}$ into a subset $\mathcal{A}$ of *atomic* individuals and the remainder set $\mathcal{C} = \mathcal{U} \setminus \mathcal{A}$ of corporate or group individuals, and assume a membership function $members : \mathcal{C} \to \mathcal{G}$. It is natural to extend $members$ to all of $\mathcal{U}$ by defining $members(i) = \emptyset$ for each atomic i. We assume that the membership hierarchy contains no cycles, and that the depth of the hierarchy is bounded by some finite number.

We do not identify group individuals with the set of their members, so two different corporate groups may happen to have the same set of members without being the same group individual. Indeed, we do not require that group individuals have any members at all. We therefore cannot use lack of members as distinguishing corporate from atomic individuals. In particular, such an extensional view of corporations would make all atomic individuals the same, as each has the same (empty) set of members.

Membership alone is a crude proxy for the great variety of different roles that individuals play in a realistic organizations, but it is all that will be considered here. Doyle [13] treats a simple case of individuals that represent others in the political sense, and Wicker [60] presents a related formalism for analyzing social change that provides for individuals standing in different relations to each other, as means to understanding the influences that groups exert on individuals and each other.

3. Belief and mental states

One cannot develop a reasonable theory of belief on the basis of sets of beliefs alone. Understanding and explaining changes of belief also requires knowing something about the structure, content, and environment of belief states. In the following, we will introduce and motivate each of these notions. The present section sets out the conception of belief and mental attitudes within which we examine

these issues (Section 3.2), the forms of ascription of belief to individuals in different states or configurations (Section 3.3), and the ways in which states of individuals relate to mental configurations (Section 3.4). The following section treats details of the primary forms of mental configurations, including logical and constitutive connections between beliefs that underlie the notion of coherence (Section 4.1), grounding of beliefs via reasons and arguments (Section 4.2), and preferences guiding choices among alternative assumptions (Section 4.3).

3.1 States

We assume that for each individual $i \in \mathcal{U}$, one identifies a set Ψ_i of all the *states* that i might inhabit at each time of interest. We assume that groups inhabit states similarly, with Ψ_g denoting the set of possible states of group $g \in G$. If $g \subseteq g'$, we assume that states of the more inclusive group determine states of the less inclusive group. We can thus describe this determination in terms of a projection mapping $\pi_\Psi(g, g') : \Psi_g \to \Psi_{g'}$. In the simplest setting, it is natural to assume that Ψ_g is isomorphic to the product $\prod_{i \in g} \Psi_i$ of the constituent individual states. Even if group states do not take this simple product form, we will assume that if $g = \{i\}$, then Ψ_g is isomorphic to Ψ_i. We will sometimes write as if $\Psi_g = \Psi_i$ in this case. We make no assumptions here about the state of a group individual, other than that a group individual typically inhabits states that need not be identical to the product state of its members.

We do not assume that the states so identified are Markovian, that is, we make no general assumption that the behavior of an individual depends only on the instantaneous state inhabited by the individual regardless of what prior states preceded the current one. Instead, we allow that changes occurring at some time might depend on past or even future states, although in the cases examined here changes will depend at most on the very recent past (first or second derivatives).

We also do not assume that the states in Ψ_i are strictly mental ones, but do not treat questions of physical embodiment in the following. If states have both mental and physical components, one can expect some degree of indeterminacy in considering mental and physical states, with distinct physical states corresponding to the same mental state, or with distinct mental states corresponding to the same physical state, depending on how one regards the relation between mind and body.

3.2 Belief content distinctions

We write $\mathbf{B}_i$ to denote a set chosen to distinguish the contents of different beliefs that one might ascribe to an individual $i \in \mathcal{U}$. For simplicity, we also will use the same set $\mathbf{B}$ across all individuals, even though differences in the conceptual apparatus of individuals, such as languages of thought, are important in contexts not treated here. We need not assume that every belief in $\mathbf{B}$ can be ascribed to each individual; indeed, we might obtain $\mathbf{B}$ as the union $\bigcup_{i \in \mathcal{U}} \mathbf{B}_i$.

We assume that $\mathbf{B}$ is finite or countably infinite and can be represented by a recursive set using a suitable grammar to express the content of beliefs. In most instances, we will proceed as though $\mathbf{B}$ is the set of sentences in a first-order predicate or modal logic.

We make no assumptions about how expressive the set of possible beliefs need be other than to assume the existence of a recursive function $\neg : \mathbf{B} \to \mathbf{B}$ that maps each belief b to an opposite belief $\neg b$. We do not assume that beliefs in $\mathbf{B}$ are individuated in semantic terms, and so allow the possibility that different beliefs can involve different statements that have the same meaning, but we also do not assume that syntactic structure never encodes any semantic structure. In particular, we allow that beliefs might embed a degree of semantic structure in which $\neg(\neg b) = b$, or even that

beliefs form the elements of a Lindenbaum algebra in which each belief constitutes an equivalence class of logically equivalent statements.

As mentioned previously, understanding realistic notions of belief change requires consideration of additional aspects of mental states besides belief. In the following, we identify roles for other types of mental attitudes, namely desires, preferences, and intentions, and for structural aspects of mind related to belief change, such as reasons or structural connections between beliefs. To provide a uniform but abstract framework for considering mental states containing these other elements as well as beliefs, we phrase the development in terms of a set $\mathcal{D}$ that contains $\mathbf{B}$ as well as other attitudes and mental elements we find cause to consider, such as sets $\mathbf{D}$ of desires and preferences, $\mathbf{I}$ of intentions, and $\mathbf{R}$ of inference policies we call reasons. In simple cases, one can regard $\mathcal{D}$ as the simple union $\mathbf{B} \cup \mathbf{D} \cup \mathbf{I} \cup \mathbf{R} \cup \ldots$ of the sets of possible beliefs, preferences, reasoning policies, arguments, and other mental attitudes. More generally, we can regard $\mathcal{D}$ as a set of mental elements or properties such that particular elements or properties can constitute more than one mental attitude.

3.3 Belief ascriptions

We regard belief ascriptions as observable properties of individuals and groups. We make no assumptions about the methods by which one might ascribe beliefs in the absence of knowledge of the specific forms of mental states examined in the following, but we do consider two possible form of such ascriptions, *binary* ascription, in which one ascribes a set of beliefs to the individual, and *graded* ascription, in which one ascribes degrees to which the individual holds each possible belief. Most of our attention in Sections 4, 5, and 6 will be on binary belief ascriptions, turning to focus on graded ascriptions in Section 7.

3.3.1 BINARY ASCRIPTION

In the *binary* form of ascription, one ascribes a set $bel_i^*(\psi) \subseteq \mathcal{D}$ to individual i. We use RMS terminology [9] to say that $b \in \mathcal{D}$ is *In* (the set of held beliefs) if $b \in bel_i^*(\psi)$, and is *Out* otherwise.

Because we assume that the set of possible beliefs is closed under negation, we obtain four possible states of binary ascriptions regarding each belief b, as depicted in Figure 1: either (1) b is *In* and $\neg b$ is *Out* the set of ascribed beliefs, (2) $\neg b$ is *In* and b is *Out*; (3) both b and $\neg b$ are *In*; or (4) both b and $\neg b$ are *Out*. The first two of these states represent the normal mode of unambiguous belief in either b or its negation. The third state represents the unfortunate condition of having conflicting or inconsistent beliefs. The fourth state represents the common condition of having no opinion on b or its negation. These four states of affairs form the basis of some relevance logics for belief first developed by Dunn [26] and Belnap [4], and for some of the bilattices studied by Ginsberg [32].

The following development describes the sets of beliefs and other mental elements of interest in two different but equivalent ways: as sets in $2^{\mathcal{D}}$ and as vectors in the binary vector space of characteristic functions over $\mathcal{D}$. Formally, we write $\mathbb{D}$ to denote the binary vector space $\mathbb{Z}_2^{\mathcal{D}}$ over $\mathbb{Z}_2$. By the natural vector-set correspondence that reads 1 as meaning that the corresponding element of $\mathcal{D}$ is *In* the set and reads 0 meaning that the element is *Out* of the set, each vector in $\mathbb{D}$ corresponds to a subset of $\mathcal{D}$.

We write $\mathbf{0} = (0, 0, \ldots)$ and $\mathbf{1} = (1, 1, \ldots)$, respectively, to denote the binary all-zeros and all-ones vectors, so that $\mathbf{0}$ represents $\emptyset$ and $\mathbf{1}$ represents $\mathcal{D}$. We define $\mathbf{1} - x$ to denote $\overline{x} = \mathcal{D} \setminus x$. Vector addition corresponds to symmetric difference, so $x + x = \mathbf{0}$ and $x - y = x + y$. Pointwise

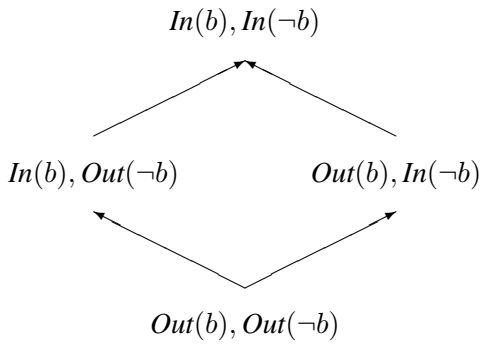

Figure 1: Fourfold states of information regarding a belief b, in which a belief and its contrary can be held or not independently of each other. The state at the head of each arrow contains more information than the state at the tail.

multiplication corresponds to intersection, so that xy corresponds to $x \cap y$. We obtain the difference $x \setminus y$ as $x\overline{y}$ and the union $x \cup y$ as the sum of products $x\overline{y} + xy + \overline{x}y$.

A binary belief ascription $bel_i^* : \Psi_i \to \mathbb{D}$, as just described, constitutes a *complete* ascription; not complete in the logical sense that every possible belief is either held or its contrary is held, but in the sense that every possible belief is either known to hold or to not hold. Although such complete ascriptions seem natural when considering ascribing beliefs to humans, they seem less natural when considering ascriptions to artificial computational individuals that build up beliefs by computing sets of beliefs in piecemeal fashion. In ascribing beliefs to such individuals, it can be useful to consider partial ascriptions that do not indicate whether a belief is held or not.

An *incomplete* binary ascription $bel_i^* : \Psi_i \to \mathbb{D} \times \mathbb{D}$ specifies *In* and *Out* labels for some proper subset of $\mathbf{B}$, and says nothing about whether the remaining possible beliefs are held or not. We characterize partial binary ascriptions in terms of pairs (x, y) in $\mathbb{D} \times \mathbb{D}$, also writing $x \setminus\setminus y$, read as "x without y", as an alternative to the pair notation. The first component x of such a pair characterizes the set of *In* beliefs, the second component y characterizes the set of *Out* beliefs, and beliefs in neither set, that is, beliefs in $\bar{x}\bar{y}$, constitute beliefs that are neither *In* nor *Out*, a status we call *Nyl*, for "not yet labeled". If $y = \overline{x}$, then $x \setminus\setminus \overline{x}$ indicates a complete binary ascription.

If $xy \neq \mathbf{0}$ in $x \setminus\setminus y$, the two sets overlap and thus constitute an inconsistent ascription. We could avoid consideration of inconsistent ascriptions by instead regarding a partial ascription as assigning one of the three labels *In*, *Out*, and *Nyl* to each element of $\mathcal{D}$. Doing so, however, complicates the algebraic structure of the set of possible ascriptions The space $\mathbb{D} \times \mathbb{D}$, like $\mathbb{D}$, is a vector space over $\mathbb{Z}_2$, but phrasing a three-valued ascription directly in vector terms would mean replacing the vector space $\mathbb{D}$ with spaces based on $\mathbb{Z}_3^{\mathcal{D}}$ or $\mathbb{D}_\perp = \{0, 1, \perp\}^{\mathcal{D}}$ over $\mathbb{Z}_2$.

One can regard the partial ascription $x \setminus\setminus y$ as denoting the interval $[x, \bar{y}]$ in the inclusion-ordered lattice of subsets of $\mathcal{D}$, namely the interval $[x, y] = \{z \subseteq \mathcal{D} \mid x \subseteq z \subseteq \bar{y}\}$. For a complete ascription, we have $[x, \bar{x}] = \{x\}$. We say a complete ascription z is compatible with or an extension of (x, y) iff $z \in [x, y]$.

3.3.2 GRADED ASCRIPTION

In the *graded* form of ascription, one ascribes to each belief a numerical grade or degree in $\mathbb{R}$. We write $\mathbb{G}$ to denote the vector space $\mathbb{R}^{\mathcal{D}}$ over $\mathbb{R}$, and write $grade_i^* : \Psi_i \rightarrow \mathbb{G}$ to denote a graded ascription function.

Most approaches to graded belief ascription focus on grades taking values in the interval $[0, 1]$ that represent probability values or fuzzy truth values, but we do not restrict attention to functions in $[0, 1]^{\mathcal{D}}$ because portions of the subsequent development require that $\mathbb{G}$ constitutes a vector space, and the functions in $[0, 1]^{\mathcal{D}}$ do not form a vector space over $\mathbb{R}$ or $\mathbb{Z}_2$ in the obvious algebras.

Graded ascriptions might be made partial in many different ways, from simple omission of grades for some beliefs, to arbitrary bits of information about individual grades themselves. The most straightforward analog of the preceding approach to partial binary ascriptions is to regard a partial graded ascription $grade_i^* : \Psi_i \rightarrow \mathbb{G} \times \mathbb{G}$ as specifying a subinterval of $\mathbb{R}$ for each potential belief, that is, a pair of functions $(x, y) \in \mathbb{G} \times \mathbb{G}$ in which we interpret $x(b)$ as providing a lower bound on the grade of b and $y(b)$ as providing an upper bound on the grade of b. In this approach, the improper subinterval $\mathbb{R}$ provides no information about the grade, while a degenerate subinterval $[\gamma, \gamma]$ ties the grade down exactly. Of course, such interval grades cannot express all forms of partial information about grades. In particular, no element of $\mathbb{G} \times \mathbb{G}$ represents the improper interval $\mathbb{R}$.

3.3.3 TRANSLATING BETWEEN ASCRIPTIONS

Summarizing the preceding, we assume that one can ascribe complete or incomplete binary or graded beliefs to states:

$$\Psi_i \left\{ \begin{array}{ll} \stackrel{bel_i^*}{\longrightarrow} & \mathbb{D} \\ \stackrel{Bel_i^*}{\longrightarrow} & \mathbb{D} \times \mathbb{D} \\ \stackrel{grade_i^*}{\longrightarrow} & \mathbb{G} \\ \stackrel{Grade_i^*}{\longrightarrow} & \mathbb{G} \times \mathbb{G} \end{array} \right. \tag{1}$$

We do not assume or require any specific relation between binary and graded ascriptions to the same individual, and in Section 3.4 regard such relations as characteristic of different types of individuals. The subsequent development will consider some translations that make sense for specific mental structures under consideration. For the moment, we merely sketch some of the most obvious possibilities and problems.

One can, for example, translate a complete binary ascription into a graded one by reinterpreting the $0, 1$ values of vectors in $\mathbb{D}$ as the $0, 1$ values of vectors in $\mathbb{G}$, and so obtain the translation

$$grade^*(\psi)(b) \quad = \quad bel^*(\psi)(b). \tag{2}$$

This same formula serves to define the translation for partial binary ascriptions into partial graded ascriptions by assigning the interval $[1, 1]$ to *In* elements, assigning $[0, 0]$ to *Out* elements, and assigning $[0, 1]$ to all *Nyl* elements. These translations need not be appropriate for conceptions of the grading space $\mathbb{G}$ different than the one considered here, and need not always be appropriate even for the grading space treated here.

The obvious means for going the other way and translating graded ascriptions into binary ones leave many more questions open. Focusing on grades that fall in the interval $[0, 1]$, one can translate a partial graded ascription into a partial binary ascription by assigning *In* to all elements graded

$[1, 1]$, assigning *Out* to all elements graded $[0, 0]$, but then one has many choices for how to label the remaining elements, the simplest being to label each of them with *Nyl*. In this case, starting with a partial binary ascription, translating to a partial graded ascription, and then back to a partial binary ascription yields the same as one started with, but a double translation starting from a partial graded ascription will generally not return to the initial ascription. But with $\mathbb{G}$ not restricting grades to the unit interval, even the preceding range of translations need not apply.

3.4 States and mental configurations

If belief ascriptions constitute observable properties of individuals, states and mental configurations characterize the underlying reality of individuals under observation. Mental configurations constitute the internal mental structures characteristic of the individual, and serve as the primary notion of mental structure that shapes belief ascription and belief change. States, in contrast, express and limit indeterminacy or uncertainty about configurations and changes of configurations of individuals.

Conceptions of mental configurations are as numerous as conceptions of mind. Most of the conceptions of mental configurations considered in the following involve more complex structures over $\mathcal{D}$ than mere binary or graded belief ascriptions. The few varieties considered here seek to understand belief change by distinguishing sets of base beliefs from conclusion beliefs, by picking out arguments supporting conclusion beliefs, and by identifing recent changes to beliefs. We write Φ_i to denote the set of mental configurations possible for individual i, and provide candidates for what such sets might contain in Section 4.

We assume that each type of mental configuration supports both binary and graded belief ascriptions, either partial or complete. That is, we require the identification of binary ascription functions $bel_i : \Phi_i \to \mathbb{D}$ or $Bel_i : \Phi_i \to \mathbb{D} \times \mathbb{D}$ and graded ascription functions $grade_i : \Phi_i \to \mathbb{G}$ or $Grade_i : \Phi_i \to \mathbb{G} \times \mathbb{G}$, which we summarize as

$$\Phi_i \begin{cases} \xrightarrow{bel_i} & \mathbb{D} \\ \xrightarrow{Bel_i} & \mathbb{D} \times \mathbb{D} \\ \xrightarrow{grade_i} & \mathbb{G} \\ \xrightarrow{Grade_i} & \mathbb{G} \times \mathbb{G} \end{cases} \tag{3}$$

It is easily seen that each complete binary or graded ascription function can be recast in the form of a partial ascription function and, as noted earlier, each binary ascription function can be recast as a graded ascription function, so these multiple requirements do not represent different constraints. Nevertheless, we will focus on binary ascriptions in the following.

In the following, we consider individuals characterized by three types of relations between states of the individual in Ψ_i and mental configurations in Φ_i, namely definite states that correspond to specific configurations in Φ_i, indefinite states that correspond to sets of configurations in 2^{Φ_i}, and distributional states that correspond to probability distributions over possible configurations in $\Pr[\Phi_i]$. Writing β_i to represent the function associating the configurational structures of individual i with states of individual i, we summarize the relations between states and configurations as follows.

$$\Psi_i \xrightarrow{\beta_i} \begin{cases} \Phi_i \\ 2^{\Phi_i} \\ \Pr[\Phi_i] \end{cases} \tag{4}$$

Complementing the function β_i that interprets states in terms of configurations, we assume ascription-combination functions $\beta^* bel_i$, $\beta^* Bel_i$ $\beta^* grade_i$, and $\beta^* Grade_i$ that ascribe beliefs to the configurational representations of states, that is, functions

$$\left.\begin{array}{l} \Phi_i \\ 2^{\Phi_i} \\ \Pr[\Phi_i] \end{array}\right\} \begin{array}{ll} \stackrel{\beta^* bel_i}{\longrightarrow} & \mathbb{D} \\ \stackrel{\beta^* Bel_i}{\longrightarrow} & \mathbb{D} \times \mathbb{D} \\ \stackrel{\beta^* grade_i}{\longrightarrow} & \mathbb{G} \\ \stackrel{\beta^* Grade_i}{\longrightarrow} & \mathbb{G} \times \mathbb{G} \end{array} \tag{5}$$

such that the beliefs ascribed to states directly match those assigned to configurational structures in the sense that

$$\begin{aligned} bel_i^*(\psi) &= \beta^* bel_i(\beta_i(\psi)) && (6) \\ Bel_i^*(\psi) &= \beta^* Bel_i(\beta_i(\psi)) && (7) \\ grade_i^*(\psi) &= \beta^* grade_i(\beta_i(\psi)) && (8) \\ Grade_i^*(\psi) &= \beta^* Grade_i(\beta_i(\psi)), && (9) \end{aligned}$$

or schematically,

$$\Psi_i \stackrel{\beta_i}{\longrightarrow} \left\{\begin{array}{l} \Phi_i \\ 2^{\Phi_i} \\ \Pr[\Phi_i] \end{array}\right\} \begin{array}{lll} \stackrel{\beta^* bel_i}{\longrightarrow} & \mathbb{D} & \stackrel{bel_i^*}{\longleftarrow} \\ \stackrel{\beta^* Bel_i}{\longrightarrow} & \mathbb{D} \times \mathbb{D} & \stackrel{Bel_i^*}{\longleftarrow} \\ \stackrel{\beta^* grade_i}{\longrightarrow} & \mathbb{G} & \stackrel{grade_i^*}{\longleftarrow} \\ \stackrel{\beta^* Grade_i}{\longrightarrow} & \mathbb{G} \times \mathbb{G} & \stackrel{Grade_i^*}{\longleftarrow} \end{array} \Psi_i. \tag{10}$$

3.4.1 DEFINITE CONFIGURATION STATES

An individual has *definite* configuration states when its states determine mental configurations uniquely, in which case the association of configurations with states takes the form of a function $\beta_i : \Psi_i \to \Phi_i$. If we seek to simplify the analysis by identifying definite states with the configurations they determine, then we have $\Psi_i = \Phi_i$ and β_i is the identify map on Φ_i. In this case we can simplify the earlier equations to obtain

$$\begin{aligned} bel_i^*(\phi) &= \beta^* bel_i(\phi) && (11) \\ Bel_i^*(\psi) &= \beta^* Bel_i(\phi) && (12) \\ grade_i^*(\psi) &= \beta^* grade_i(\phi) && (13) \\ Grade_i^*(\psi) &= \beta^* Grade_i(\phi) && (14) \end{aligned}$$

3.4.2 INDEFINITE CONFIGURATION STATES

An individual has *indefinite* configuration states when its states determine a set of possible mental configurations, in which case the association of configurations with states takes the form of a function $\beta_i : \Psi_i \to 2^{\Phi_i}$ that maps each state $\psi \in \Psi_i$ to a set $\beta_i(\psi)$ of belief states in Φ_i. We leave open the possibility that $\beta_i(\psi)$ can be empty, finite, or infinite. One can regard definite configuration states as a special case of indefinite configuration states, in which each state determines a singleton set of mental configurations.

Nonmonotonic logics and other approaches involving grounded belief configurations such as those treated in Section 4.2 provide the most common motivation for considering indefinite belief individuals. The simplest conception of grounded belief configuration regards each mental configuration $\phi \in \Phi$ as a pair $(m, x) \in \mathbb{D} \times \mathbb{D}$ in which m consists of a set of base elements in $\mathcal{D}$, and x consists of a set of conclusion elements in $\mathcal{D}$. One typically can have different sets of conclusions grounded in the same set of base beliefs. In this setting, one assumes that each mental configuration determines a unique set of base beliefs, and that the set of possible belief states consist of the set of "expansions" of the belief base.

Unlike the case of definite configuration states, in which we can obtain ascriptions to states directly from ascriptions to configurations, many alternatives exist for constructing or defining ascriptions to states from ascriptions to sets of configurations. Indeed, we regard the relation between these types of ascriptions, as expressed in the functions $\beta^* bel_i$, $\beta^* Bel_i$, $\beta^* grade_i$, and $\beta^* Grade_i$ as characteristic of the individual, as part of its constitution, so that one finds different relations characteristic of different individuals. We will identify some such relations characteristic of special types of believing individuals in Section 7, but otherwise leave the question open.

Consider first indefinite ascriptions based on complete binary ascriptions. Borrowing terminology from [41] introduced in the context of grounded belief, we say that b is *arguable* iff there is some $\phi \in \beta_i(\psi)$ such that $b \in bel_i(\phi)$ and that b is *provable* iff $b \in bel_i(\phi)$ for each $\phi \in \beta_i(\psi)$. We also say that b is *doubtless* iff there is some $\phi \in \beta_i(\psi)$ such that $\neg b \notin bel_i(\phi)$ and that b is *conceivable* iff $\neg b \notin bel_i(\phi)$ for each $\phi \in \beta_i(\psi)$. In definite belief individuals, b is *In* iff b is arguable iff b is provable. We can depict the inclusions between these classes of beliefs as in Figure 2.

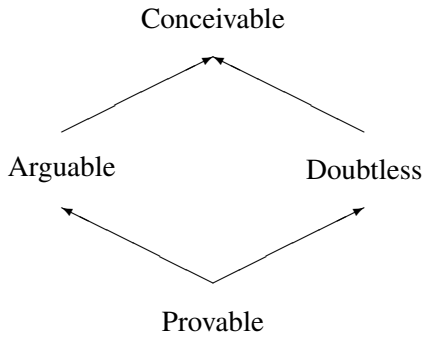

Figure 2: Inclusion relationships between classes of beliefs distinguished by the appearance of a belief and its contrary in alternative mental configurations. The class at the head of each arrow includes the class at the tail.

For the case of partial binary ascriptions, one gets a more complicated set of descriptors and interrelations. These four descriptors still apply because their definitions refer only to what is *In*. One gets additional concepts by considering what is *Out* and *Nyl*. One obtains a still more complicated picture when considering complete and partial graded ascriptions.

3.4.3 DISTRIBUTIONAL CONFIGURATION STATES

An individual has *distributional* configuration states when its states determine probability distributions over all possible mental configurations, in which case the association of configurations with states takes the form of a function $\beta_i : \Psi_i \rightarrow \Pr[\Phi_i]$, where $\Pr[\Phi_i]$ denotes the set of all probability distributions over mental configurations in Φ_i. Definite belief individuals form a special case of distributional belief individuals, in which each state assigns probability 1 to one configuration and probability 0 to all others.

We do not attempt to examine here all possibilities for defining $\beta^* bel_i$, $\beta^* Bel_i$, $\beta^* grade_i$, and $\beta^* Grade_i$ for distributional individuals. There are 16 cases to consider, namely complete or incomplete binary or graded ascriptions to states derived from complete or incomplete binary or graded ascriptions to configurations. More importantly, different types of individuals might reasonably exhibit different constitutions, in which case there is no single definition for all cases. At this point we only illustrate two of the possibilities for defining ascriptions to states based on ascriptions to configurations, and return to this question later to consider additional possibilities.

First, one might derive complete graded ascriptions from complete binary ascriptions as follows. When combined with a complete binary belief ascription, each probability distribution $Pr_\Phi : \Phi \rightarrow [0, 1]$ over mental configurations induces a measure $\beta^* grade : \mathcal{D} \rightarrow \mathbb{R}$ over elements in $\mathcal{D}$ by the definition

$$\beta^* grade(b) = \int Pr_\Phi(\{\phi \in \Phi \mid b \in bel(\phi)\}), \tag{15}$$

which we abbreviate by writing $Pr(b)$ to mean $\beta^* grade(b)$. These induced measures are not in general probability measures in the absence of assumptions about the completeness and consistency of the possible belief ascriptions, for without such assumptions, one can have $Pr(b) + Pr(\neg b) < 1$ or $Pr(b) + Pr(\neg b) > 1$. In particular, for definite belief individuals, $Pr(b) = 1$ if b is ***In***, $Pr(b) = 0$ if b is ***Out***, and $Pr(b) + Pr(\neg b)$ can take the values 0, 1, or 2.

Second, one might derive incomplete graded ascriptions to states from incomplete binary ascriptions to configurations as follows. In this case, we derived $\beta^* grade : \mathcal{D} \rightarrow (\mathbb{R} \times \mathbb{R})$ by

$$\beta^* grade(b) = \left(\int Pr_\Phi(\{x \setminus\setminus y \in \Phi \mid b \in x\}), \int Pr_\Phi(\{x \setminus\setminus y \in \Phi \mid b \notin y\}) \right), \tag{16}$$

in which the lower bound in this confidence interval is the probability of only the beliefs in x being ***In***, and the upper bound is the probability of only the beliefs in y being ***Out***.

4. The structure of belief configurations

We distinguish three dimensions along which approaches to the study of belief change differ. These dimensions represent different assumptions about the properties required of mental configurations. We do not claim these represent the only ways of distinguishing approaches, only that they represent distinctions clearly visible in current approaches.

- *Coherence:* Rather than follow traditional epistemic logics in assuming that beliefs cohere in the sense of exhibiting ideal logical closure and consistency, mental configurations can exhibit coherence based on weaker constitutive logics that requires beliefs to exhibit lesser degrees of consistency and closure, and that coherence might involve nonlogical connections, including nonmonotonic closure conditions.

- *Foundation:* Rather than consisting of a set of beliefs related only by coherence criteria, beliefs might admit a division into base beliefs and those beliefs derived or constructed from the basis beliefs.

- *Optimality:* Rather than only depending on beliefs, mental configurations might represent the result of choices among alternatives based on preferences among beliefs and mental configurations.

The following will present representatives of some alternatives along each of these dimensions in some detail.

4.1 Coherence of beliefs

Quine painted a picture of mental change in his image of the "web of belief" [48, 47], in which changes of belief are restrained or shaped by a network of connections between beliefs, such that the overall change resulting from some sets of pokes and tugs satisfies his principle of *minimum mutilation* [46], in which the resultant change is the "smallest" change that accomodates the perturbing pokes and tugs.

Philosophy and artificial intelligence have explored connections between beliefs of several types, including notions of logical entailment and consistency that connect beliefs through their meanings, and notions of nonmonotonic and nonlogical inference that connect beliefs through what one might think of as habits of thought or patterns of plausible reasoning. These types of connection between beliefs certainly do not exhaust the ideas relevant to understanding belief change. For example, logical and reasoning dependencies between beliefs do not touch on temporal connections between beliefs, the most prominent of which is the persistence of memory, in which a belief concluded at some point in the past is believed currently only because nothing worked to remove it. Nor do logical and reasoning dependencies reflect certain kinds of active dependence in which motives maintain beliefs by constant action that resists change and restores their targets. We consider these types of connections between beliefs in subsequent sections.

Formalizing notions of coherence requires only a very simple conception of mental configuration, namely as a complete or partial set of beliefs and other mental elements, that is, configurations of type $\mathbb{D}$ or of type $\mathbb{D}^2$. The belief ascription functions for each of these is obviously the identity, that is, $bel(\phi) = \phi$ for $\phi \in \mathbb{D}$ and $Bel(\phi) = \phi$ for $\phi \in \mathbb{D}^2$, the former being complete ascriptions and the latter partial.

Of course, not every subset need be considered coherent, so we identify assumptions about the mental constitution [13, 19] that can (but need not) restrict coherent configurations to including only some elements of $\mathbb{D}$ or $\mathbb{D}^2$.

4.1.1 CONSISTENCY AND CLOSURE

This section considers constitutive assumptions requiring that beliefs ascribed to mental configurations must satisfy logical consistency and closure conditions. For example, one type of constitution might rule out overtly inconsistent sets of beliefs by restricting the individual to beliefs that never include both b and $\neg b$; another might require that the beliefs contain b and c if it contains a belief expressing "b and c."

We treat constitutive logics abstractly by using Scott [53] *information systems* to present the requirements of inherent inference in direct terms rather than in terms of axioms phrased in some

particular logical language. A Scott information system over $\mathcal{D}$ is chararacterized formally by three items $(\mathcal{D}, \mathit{Con}, \vdash)$, in which $\mathcal{D}$ constitutes the set of atomic "propositions". The set *Con* consists of the finite subsets of $\mathcal{D}$ that are "consistent", that is, that satisfy the constitutive consistency condition, and the relation $\vdash$ on $\mathit{Con} \times \mathcal{D}$ consists of the finite "entailments" inherent in the constitutive inference conditions, writing $x \vdash a$ to mean (x, a) is in $\vdash$. To capture what we would think of as actual consistency and entailment conditions, *Con* and $\vdash$ must satisfy the following conditions.

The conditions on consistency are that subsets of consistent sets are consistent, or formally, that $x \in \mathit{Con}$ if $x \subseteq y \in \mathit{Con}$; and that each element is itself consistent, or formally, that $\{a\} \in \mathit{Con}$ if $a \in \mathcal{D}$. We say that a set $x \subseteq \mathcal{D}$ is *consistent* iff each finite subset $y \subseteq x$ is consistent according to *Con*.

The conditions on entailment state that addition of entailed elements preserves consistency, or formally, that $x \cup \{a\} \in \mathit{Con}$ if $x \vdash a$; that consistent sets entail their own members, or formally, that $x \vdash a$ if $a \in x$; and that entailment is transitive, or formally, that $y \vdash a$ if $y \vdash b$ for all $b \in x$, and $x \vdash a$. We say that x is *(inferentially) closed* iff $a \in x$ whenever $y \subseteq x$ and $y \vdash a$. We extend the notation of entailment in the natural way to say that $x \vdash y$ iff $x \vdash a$ for each $a \in y$, in which case the transitivity of entailment condition can be rewritten as $x \vdash Z$ whenever $x \vdash y$ and $y \vdash Z$. We then introduce a *closure* operator θ, defining $\theta(x)$ to be the least closed superset of x. We also broaden the notation $x \vdash a$ to mean $a \in \theta(x)$.

To apply the notion of inherent logic to mental configurations of type $\mathbb{D}$ or type $\mathbb{D}^2$, we require that the set *In* elements ascribed to the configuration be consistent and closed. For configurations of type $\mathbb{D}$, this means the set $x \in \mathbb{D}$, and for configurations of type $\mathbb{D}^2$, this means the first component of $(x, y) \in \mathbb{D}^2$.

These abstract notions of consistency and closure cover many possibilities, ranging from the ordinary logical notions of ideal consistency and entailment to individuals who enjoy no nontrivial powers to identify inconsistencies or entailments, for which case one considers every finite set to be consistent, and for entailment to be just simple containment $x \vdash a$ iff $a \in x$, so that every subset of $\mathcal{D}$ is both consistent and closed. These notions also provide the means to require that sets of conclusions always contain certain elements, in that an entailment $\emptyset \vdash a$ requires that a be in every closed set.

4.1.2 CONNECTIONS BETWEEN BELIEFS

Logical consistency and closure notions do not exhaust the properties relevant to coherence of beliefs. In particular, consistency and closure are incapable of expressing nonmonotonic connections between beliefs, such as a requirement that b should be held if $\neg b$ is not.

In the following, we consider coherence constraints that can be expressed as connections between specific beliefs, which we interpret formally as the requirement that one can express each such constraint as a specific Boolean combination of *In* and *Out* predicates, such as $\mathit{In}(b) \vee \mathit{In}(\neg b)$. The original motivation for considering such cognitive connections was as means for recording finite traces of past derivations in a reason maintenance system (RMS) [9]. In this setting, these recorded *reasons* represent logical or nonlogical inferential relationships among beliefs and other mental attitudes and representations. One may also regard them as mental habits or policies regarding when to hold or not hold certain beliefs, or as nonmonotonic closure policies or constraints on sets of conclusions. As such connections play important roles in understanding commonsense

Reason type	**Full form**	**Abbreviation**
Premises	$\emptyset \setminus\setminus \emptyset \mid\!\!\!\vdash C \setminus\setminus \emptyset$	$\mid\!\!\!\vdash C$
Monotonic reasons	$A \setminus\setminus \emptyset \mid\!\!\!\vdash C \setminus\setminus \emptyset$	$A \mid\!\!\!\vdash C$
Nonmonotonic defaults	$A \setminus\setminus B \mid\!\!\!\vdash C \setminus\setminus \emptyset$	$A \setminus\setminus B \mid\!\!\!\vdash C$
Normal defaults	$A \setminus\setminus \{\neg b\} \mid\!\!\!\vdash \{b\} \setminus\setminus \emptyset$	$A \setminus\setminus \{\neg b\} \mid\!\!\!\vdash \{b\}$
Pure assumptions	$\emptyset \setminus\setminus \{\neg b\} \mid\!\!\!\vdash \{b\} \setminus\setminus \emptyset$	$\setminus\setminus \{\neg b\} \mid\!\!\!\vdash \{b\}$
Pure denials	$\emptyset \setminus\setminus \emptyset \mid\!\!\!\vdash \emptyset \setminus\setminus D$	$\mid\!\!\!\vdash \setminus\setminus D$

Figure 3: Special cases of interval conditionals and their abbreviations.

reasoning, a comprehensive theory of coherence must augment the logically monotonic notions of consistency and closure with means for characterizing nonmonotonic constraints between beliefs.

In line with these motivations, we focus on reasons that can be phrased as conditionals written in the form $A \setminus\setminus B \mid\!\!\!\vdash C \setminus\setminus D$, for $A, B, C, D \subseteq \mathcal{D}$, read as "*antecedents* A without *defeaters* B gives *consequences* C without *denials* D," meaning that each element of the consequences C must be *In* and each element of the exclusions D must be *Out* if each element of A is *In* and each element of B is *Out*. For example, one could cast the requirement "Hold (c) 'Sasha can fly' if (a) 'Sasha is a bird' is not held, and (b) 'Sasha cannot fly' is held" as the reason $\{a\} \setminus\setminus \{b\} \mid\!\!\!\vdash \{c\} \setminus\setminus \{\}$.

Formally, each such reason corresponds to a coherence condition on mental configurations. We write $[\![A \setminus\setminus B \mid\!\!\!\vdash C \setminus\setminus D]\!]$ to denote the meaning of this condition, namely the set of mental configurations satisfying the condition. For configurations of type $\mathbb{D}$, the satisfying configurations are given by

$$[\![A \setminus\setminus B \mid\!\!\!\vdash C \setminus\setminus D]\!] = \{x \in \mathbb{D} \mid A\overline{x} + Bx = \mathbf{0} \rightarrow C\overline{x} + Dx = \mathbf{0}\}, \tag{17}$$

and for configurations of type $\mathbb{D}^2$, the configurations satisfying a reason are given by

$$[\![A \setminus\setminus B \mid\!\!\!\vdash C \setminus\setminus D]\!] = \{(x, y) \in \mathbb{D}^2 \mid A\overline{x} + B\overline{y} = \mathbf{0} \rightarrow C\overline{x} + D\overline{y} = \mathbf{0}\}. \tag{18}$$

We say a reason $A \setminus\setminus B \mid\!\!\!\vdash C \setminus\setminus D$ is *valid* with respect to configurations x or (x, y) if its antecedent conditions hold, that is, if $A\overline{x} + Bx = \mathbf{0}$ or $A\overline{x} + B\overline{y} = \mathbf{0}$.

Figure 3 lists several special cases of such conditional reasons corresponding to common uses in automated reasoning systems. Premise reasons can be used to stipulate axioms or other required conclusions. Records of ordinary logical inferences take the form of monotonic reasons, but monotonic reasons in general need not be logically sound. Nonmonotonic default reasons state conditional assumptions and commonly represent heuristic or plausible inferences by making assumptions absent information that would contradict or defeat the assumptions. Nonmonotonic default reasons cannot forbid the presence of elements in mental states. Normal defaults represent a common form of simple nonmonotonic assumptions, and pure assumptions represent the unconditional form of normal defaults. Pure denials are used to mark contradictions in the dependency-directed backtracking methods of Sussman and Stallman [56, 55] and [9].

With our earlier notice that $A \setminus\setminus B$ identifies an interval in the lattice of subsets of $\mathcal{D}$, namely those subsets that contain all elements of A but contain no elements of B, we can also call reasons of the form $A \setminus\setminus B \mid\!\!\!\vdash C \setminus\setminus D$ *interval conditionals*, as such reasons require that beliefs fall in the interval $C \setminus\setminus D$ if they fall in the interval $A \setminus\setminus B$. One can transform arbitrary Boolean conditions on specific beliefs to sets of relations between intervals by conversion to disjunctive normal form

and then regarding each disjunct as either a positive interval conditional $A \setminus\setminus B \mid\!\!\sim C \setminus\setminus D$ or as a negative interval conditional $A \setminus\setminus B \mid\!\!\sim \neg(C \setminus\setminus D)$. Positive interval conditionals have the conceptual advantage of relating to individual beliefs either in a positive way (appearance in C) or a negative way (appearance in D), as opposed to general conditions in which the relation need not be either purely positive or negative. Positive interval conditionals also express definite constraints on beliefs, in contrast to negative interval conditionals, which express possibly ambiguous disjunctive constraints.

Consistency and closure represent global coherence conditions not tied to individual beliefs. If we also regard the set of cognitive connections as static, we obtain a view of belief akin to that taken in Reiter's [49] default logic, which augments ordinary logical closure and consistency with a fixed set of *defaults* that correspond to interval conditional schema without denials. In that setting, the default and logical inference combine to yield sets of conclusions via grounding definitions akin to those described in Section 4.2. In considering active reasoners, however, one finds attitudes that are naturally regarded as intentions or habits that state policies that constrain the structure of mental states and the conduct of reasoning, such as nonmonotonic reasoning rules discussed presently. Such policies can be themselves adopted or abandoned in the course of reasoning, unlike global notions of logical consistency and closure. In such cases, it is natural to regard an interval conditional as a mutable element of the mental state, as an element $r \in \mathcal{D}$ that constrains states to fall in the set $[\![r]\!] = [\![A_r \setminus\setminus B_r \mid\!\!\sim C_r \setminus\setminus D_r]\!]$ for some sets $A_r, B_r, C_r, D_r \subseteq \mathcal{D}$. We can extend this interpretation to all state elements, even ones that are not intended to constitute reasons, by interpreting non-conditional elements as the vacuous conditional $\emptyset \setminus\setminus \emptyset \mid\!\!\sim \emptyset \setminus\setminus \emptyset$, which has the unconstraining meaning $[\![\emptyset \setminus\setminus \emptyset \mid\!\!\sim \emptyset \setminus\setminus \emptyset]\!] = \mathbb{D}$ for configurations of type $\mathbb{D}$ and the meaning $\mathbb{D}^2$ for configurations of type $\mathbb{D}^2$.

With identifications of the (possibly trivial) connective meaning of every element of $\mathcal{D}$, we can formulate an expanded notion of coherence as follows.

We say that a configuration $x \in \mathbb{D}$ is *self-satisfying* just in case it satisfies the conditions expressed by each of its *In* elements, that is, just in case $x \in [\![e]\!]$ for each $e \in x$. We extend the meaning function to a function over sets of reasons by defining $[\![\emptyset]\!] = \mathbb{D}$ and $[\![x]\!] = \bigcap_{e \in x} [\![e]\!]$ for each nonzero $x \in \mathbb{D}$, so that x is self-satisfying iff $x \in [\![x]\!]$. A configuration $(x, y) \in \mathbb{D}^2$, in turn, is *self-satisfying* just in case it satisfies the conditions expressed by each of its *In* elements, that is, just in case $(x, y) \in [\![e]\!]$ for each $e \in x$. We extend this meaning function as well to a function over sets of reasons by defining $[\![\emptyset]\!] = \mathbb{D}^2$ and $[\![x]\!] = \bigcap_{e \in x} [\![e]\!]$ for each nonzero $x \in \mathbb{D}$, so that (x, y) is self-satisfying iff $(x, y) \in [\![x]\!]$. In either case, therefore, a configuration is self-satisfying just in case it respects the consequences and denials of each of its valid *In* reasons.

Summing up, we obtain a more comprehensive notion of coherence of beliefs by augmenting satisfaction of global consistency and closure conditions with satisfaction of all connective conditionals present in the state. We say that configurations x of type $\mathbb{D}$ or (x, y) of type $\mathbb{D}^2$ is *coherent* just in case its *In* elements are consistent and closed, that is, that $\theta(x) = x$, and that the configuration as a whole is self-satisfying, that is, that $x \in [\![x]\!]$. The subset of coherent vectors in $\mathbb{D}$ or $\mathbb{D}^2$ need not form a vector space.

One can exploit such mutable reasons to avoid the need to remove reasons from mental states, allowing reasoning to purely accumulate reasons over time. To do this, one assumes the existence of "defeater" elements in $\mathcal{D}$ specific to each reason and then includes these defeater elements in the constraints expressed by the reason. That is, one assumes that the element *defeated*(r) represents

the defeat of the conditional $A \setminus\setminus B \mid\!\!\sim C \setminus\setminus D$, which one then interprets as

$$r = A \setminus\setminus B \cup \{\mathit{defeated}(r)\} \mid\!\!\sim C \setminus\setminus D. \tag{19}$$

4.2 Grounding of beliefs

Gärdenfors [29, 30] distinguishes two types of theories of belief, those in which the set of beliefs is derived from some set of stipulated foundational or basis beliefs, and those in which the set of beliefs subsists on its own without any notion of foundation. Both types of theories regard beliefs as consistent and closed with respect to some logic of belief. In the ungrounded or *coherence* theories the consistency and closure constraints are the only assumptions made about the set of beliefs, while in the grounded or *foundations* theories the set of beliefs must satisfy additional constraints regarding derivation or construction from the stipulated foundation or basis [14]. For example, one might regard the basis to be some arbitrary set of beliefs, and the definite beliefs resulting to be the closure of this set under entailment in a classical logic, such as the modal logic of belief **K4**. If the logic of belief is a nonmonotonic one, such as Moore's [42] autoepistemic logic, the same base beliefs can generate a set of possible belief configurationss, corresponding to an indefinite belief state.

Our exemplar for foundational systems is that of truth or reason maintenance systems (TMS/RMS) [9], in which the base structures are called justifications or reasons, and in which reasoning is organized to systematically add reasons that represent steps of reasoning as they are made, for example, through chunking or similar notions. To obtain a belief set that contains some specific belief, one adds a new reason to the RMS base that draws the target belief as a conclusion. RMS reasons persist in memory until explicitly expunged, so the new conclusion persists as well barring further changes to the set of base reasons. To remove an existing belief, one changes the base so that it no longer draws the target belief as a conclusion. This could be done by finding and removing some base reason underlying the target belief, but RMS reasons are nonmonotonic, so that one can defeat some conclusion by adding new reasons that defeat the reasons formerly supporting the target. Removing some conclusion indirectly can require defeating more than one underlying reason, as RMS retains consequences that have alternate support.

The notion of constructive belief applies even with different notions of base beliefs. For example, one might take the base structure to consist of Bayesian networks or artificial neural networks and regard constructive beliefs as derived by means of probabilistic or neural-element thresholds. For simplicity, we focus on constructive schemes starting and ending with the elements of $\mathcal{D}$ introduced in the preceding.

4.2.1 GROUNDING CONFIGURATIONS

Standard theories of nonmonotonic reasoning focus on the possible states of conclusions derivable from base beliefs and nonmonotonic inference rules, calling these sets *extensions* or "expansions" or "answer sets". We treat such extensions presently, but develop them by considering first the notion of *grounding configurations* that identify the different ways in which conclusions can be grounded in base beliefs.

Grounding configurations correspond more closely to practicable means for producing or revising belief states via reasoning or computation from base beliefs than do the conceptions of configuration-free extensions. Rather than focus only on the conclusions reached, grounding configurations focus on the means by which these conclusions are reached. Grounding configurations thus

distinguish a set of conclusions reached by one set of arguments from the same set of conclusions reached by a different set of arguments. In this setting, a new set of beliefs constructed following a change to base beliefs might, in some cases, be the same as the previous set of constructive beliefs, in which case we regard the two belief states as having both different base beliefs and different internal configurations.

Distinguishing configurations from conclusions yields a formalization closer to the structure of computational systems like RMS than to the structure of a plain nonmonotonic logic. Configurational structure played a key role in the original RMS implementations, with a belief state consisting of both a labeling of conclusions as *In* and *Out* and an assignment of supports, such as supporting reasons, to each conclusion labeled *In*.

Writing $\mathbb{S}$ to denote the space of functions $(\mathcal{D} \to \mathbb{D})$, we define a complete belief configuration to be a triple $(m, s, x) \in \mathbb{D} \times \mathbb{S} \times \mathbb{D}$, in which we regard m as the basis set, x as the constructed set, and s as picking out a set of supports for each mental element. Naturally, a complete belief ascription of a configuration of this type consists of the third component, that is, $bel((m, s, x)) = x$. We define a partial belief configuration similarly to be a quadruple $(m, s, x, y) \in \mathbb{D} \times \mathbb{S} \times \mathbb{D}^2$, in which the components x and y characterize the *In* and *Out* conclusions separately, so that we have the natural partial belief ascription given by $Bel((m, s, x, y)) = (x, y)$.

The notion of coherent configuration described previously reflects one half of the basic RMS stability principle by requiring that state elements be *In* (or *Out*) if a valid reason requires them to be so. The simplest notion of grounding reflects the other half of the stability principle, that state elements be *In* only if they have a valid reason for being held. The notion of *local grounding* places no restriction on the valid reasons used to justify holding a belief, whereas the notion of *global grounding* requires that the justifying reasons form part of a noncircular argument for holding the belief. We treat these grounding varieties in turn, and define both only in terms of configurations of type $\mathbb{D} \times \mathbb{S} \times \mathbb{D}^2$, as complete grounding configurations correspond to a subset of these satisfying the same definitions under the constraint that $y = \overline{x}$.

4.2.2 LOCAL GROUNDING

We say that a configuration (m, s, x, y) is *locally grounded*, or that s locally grounds (x, y) in m, if the presence or absence of conclusions in (x, y) exactly corresponds to the existence of supporting reasons in s that are valid with respect to (x, y) and m. Formally, (m, s, x, y) is locally grounded iff x is coherent (consistent, closed, and self-satisfying), and for each $e \in \mathcal{D}$, $x(e) = 1$ (that is, e is *In*) iff

1. $\emptyset \subseteq s(e) \subseteq m \cup (x \setminus \{e\})$ and $s(e) \vdash e$, or

2. $\emptyset \subset s(e) \subseteq m \cup x$ and for each $r \in s(e)$, there are A, B, C, D such that $r = A \setminus\setminus B \mid\!\!\sim C \setminus\setminus D$, $A\overline{x} + B\overline{y} = \mathbf{0}$, and $e \in C$,

The first of these conditions requires that every element of m be in x because $\{e\}$ is by assumption consistent and entails its own element. If we say e is *stipulated* in x with respect to m just in case $e \in m$, then a state element is locally grounded just in case it is either stipulated, logically derivable from other *In* elements, or is a required conclusion of some *In* reason.

Mutually supporting beliefs appear frequently when reasoning about equality or logically equivalent statements. These give rise to alternative configurations for the same set of conclusions and to

self-supporting cycles in locally grounded extensions. Figure 4 displays a situation in which three base reasons support two beliefs in three distinct locally grounded configurations.

	Base reasons	Conclusions
	$\Vdash a$	a
+	$a \Vdash c$	a, c
+	$c \Vdash a$	a, c

Conclusion	Supports
c	$a \Vdash c$
	and
a	$\Vdash a$; or
	$c \Vdash a$; or
	$\Vdash a, c \Vdash a$

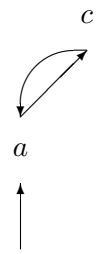

Figure 4: Three distinct configurations with same conclusions.

Locally grounded configurations permit self-supporting cycles, mutually supporting beliefs that arise easily in reasoning about equality or logically equivalent statements. For example, starting from any two of the equalities $X = Y$, $Y = Z$, and $X = Z$, one might infer the other. A simple example is displayed in Figure 5, in which one starts with no base beliefs and adds and removes various reasons, some of which connect beliefs, and others of which stipulate the presence of beliefs. At the end, one has left only the connective reasons $a \Vdash c$, $b \Vdash c$, and $c \Vdash b$. This set of reasons plays into two distinct locally grounded configurations, one in which both b and c are held, and one in which neither is held. The latter is also a globally grounded configuration, as discussed presently.

	Base reasons	Conclusions
	$\Vdash a$	a
+	$a \Vdash c$	a, c
-	$\Vdash a$	
+	$\Vdash a$	a, c
+	$b \Vdash c$	a, c
+	$\Vdash b$	a, b, c
-	$\Vdash a$	b, c
+	$c \Vdash b$	b, c
-	$\Vdash b$	b, c (local); none (global)

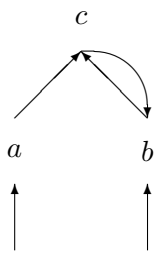

Figure 5: Local versus global grounding of conclusions. The conclusion sets are obtained by starting with an empty set of reasons and adding and removing reasons.

4.2.3 GLOBAL GROUNDING

The conception of global groundedness underlies the belief revision effected by RMS and the notions of extension and answer sets of most systems of nonmonotonic or default logics and answer-

set programming systems. In contrast to the extensions admissible in locally-grounded beliefs, grounded stable extensions omit self-supporting cycles of the sort seen in Figure 5.

We say that a configuration (m, s, x, y) is *globally grounded* if the presence or absence of conclusions in x exactly corresponds to the existence of noncircular supporting arguments starting from m. Formally, (m, s, x, y) is globally grounded iff (m, s, x, y) is locally grounded and there is a well-ordering $<$ on $\mathcal{D}$ such that for each $e \in \mathcal{D}$, either

1. $\emptyset \subseteq s(e) \subseteq m \cup x_{<e}$ and $s(e) \vdash e$, where $x_{<e} \stackrel{\text{def}}{=} \{b \in x \mid b < e\}$, or
2. $\emptyset \subset s(e) \subseteq m \cup x_{<e}$ and for each $r \in s(e)$, there are $A, B \subseteq x_{<e}$ such that $r = A \setminus\setminus B \mid\!\!\sim C \setminus\setminus D$, $A\overline{x} + Bx = \mathbf{0}$, and $e \in C$.

These conditions parallel those defining locally grounded configurations, but require that the supporting inferences or reasons be developed prior to the conclusion they support. Globally grounded extensions thus omit self-supporting cycles of the sort seen in Figure 5, requiring instead at least one noncircular derivation from base beliefs by either logic or reasons.

We say that a locally or globally grounded configuration (m, s, x, y) is *minimal* if elements obtained by entailment have minimal supporting sets, that is, if $s' \not\vdash e$ whenever $s(e) \vdash e$ and $s' \subset s(e)$. We say that the configuration is *functional* just in case the support function s identifies exactly one supporting reason for every element in x, that is, if for each $e \in x$, if $s(e) \not\vdash e$, then $|s(e)| = 1$. The original RMS [9] and related systems relied on functional globally grounded states, in which one of the possible valid reasons for a belief was distinguished as the supporting reason.

4.2.4 GROUNDED EXTENSIONS

For each locally or globally grounded configuration, we say that the complete or incomplete sets of conclusions of the configuration constitute a grounded *extension* of the base beliefs of the configuration. Thus x would be a grounded extension of m given a grounded configuration (m, s, x), and $x \setminus\setminus y$ would be a grounded extension of m given a grounded configuration (m, s, x, y). We write $\alpha^{\mathrm{L}}(m)$ to denote the set of all locally grounded extensions of m, and $\alpha^{\mathrm{G}}(m)$ to denote the set of all globally grounded extensions of m.

When all reasons are monotonic, there is one and only one globally grounded extension. There may be more than one locally grounded extensions, but in this case there is one locally grounded extension that contains all the others as proper subsets.

A different situation prevails when some reasons are nonmonotonic. Nonmonotonic logics characteristically exhibit multiple extensions in which a set of base reasons can support zero, one, or more grounded extensions, none of which need include the others. Figure 6 displays a simple case in which the full set of reasons supports an extension in which a is believed but b is not, and another in which b is believed and a is not. Moreover, reasoning exhibits hysteresis, in that adding a reason that defeats one extension but is later removed can lead to an extension other than the one originally entertained. As the accompanying table indicates, adding and removing base reasons one at a time can determine which of the extensions one has in hand. The first extension $\{a, e\}$ found will persist under such sequential revision until defeated, at which time the other extension $\{b, d\}$ will be held, even if the the reason that originally defeated it is restored.

Because beliefs might be in one extension but not in others, we require additional terms to characterize the relation of some $b \in \mathcal{D}$ to the grounded configurations of the individual. We reuse the terms presented in Figure 2 to classify beliefs as arguable, provable, doubtless, and conceivable

	Base reasons	Global conclusions
	$\backslash\backslash\, d \parallel\!\!\vdash a$	a
+	$\backslash\backslash\, e \parallel\!\!\vdash b$	a, b
+	$a, b \parallel\!\!\vdash c$	a, b, c
+	$a \parallel\!\!\vdash e$	a, e
+	$b \parallel\!\!\vdash d$	a, e
-	$a \parallel\!\!\vdash e$	b, d
+	$a \parallel\!\!\vdash e$	b, d

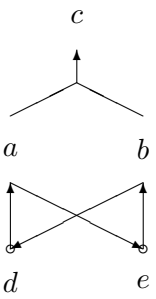

Figure 6: Two incompatible extensions, with sequential revision exhibiting hysteresis

depending on the presence or absence of a belief and its contrary in the beliefs ascribed to the extensions in $\alpha(m)$. When the logic of belief includes first-order logic, all these inclusions shown in Figure 2 are proper.

Dung [25, 5] develops a theory of abstract arguments and "attack" relationships between them that provides a general framework for formalization of multiple extensions grounded in basis beliefs.

Local and global grounding need not exhaust the possible notions of grounding. One can consider intermediate forms in which beliefs are grouped into different "regions" and consider regional grounding in which grounding is acyclic within each region but may be cyclic among regions, as in the rational distributed reason maintenance system of [20]. That treatment had in mind regions representing "subagencies" of a single mind, but one might use the same ideas in considering beliefs that are regionally grounded within subtheories corresponding to different subject matters (religious, moral, political, scientific, etc.) or authorities.

4.3 Preference among beliefs

Some theories of belief augment coherence or grounding conditions with optimality conditions, such as Rescher's *preferred maximal consistent subsets* of conflicting beliefs [50], that require that belief configurations be the best of those available, according to some standard of comparison.

Abstractly, theories of preferred configurations posit a weak preference relation $\precsim$ on mental configurations, that is, a complete ($\phi \precsim \phi'$ or $\phi' \precsim \phi$) and transitive (if $\phi \precsim \phi'$ and $\phi' \precsim \phi''$, then $\phi \precsim \phi''$) ordering of mental configurations, for which we define indifference $\sim$ by $\phi \sim \phi'$ iff $\phi \precsim \phi'$ and $\phi' \precsim \phi$, and define strict preference $\prec$ by $\phi \prec \phi'$ iff $\phi \precsim \phi'$ and $\phi' \not\precsim \phi$.

If Φ' is a set of mental configurations, we say that a configuration $\phi \in \Phi'$ is *optimal* or *maximally preferred* with respect to Φ' and $\prec$ iff there is no $\phi' \in \Phi'$ such that $\phi' \prec \phi$.

To add a requirement of optimality of indefinite belief states to requirements of coherence and grounding, we require that each indefinite belief state contains only configurations optimal with respect to the indefinite belief state. To add an optimality requirement to distributional belief states, we require that each of the configurations assigned nonzero probability by a distribution is optimal with respect to all other configurations assigned nonzero probability. For grounded definite belief states, optimality requires that each selected mental configuration be optimal with respect to the possible grounded alternatives. For nongrounded definite belief states, optimality requirements do not apply until we consider belief change, when we can require optimality of the revised mental configuration with respect to the alternative possible revisions.

Adding the requirement of optimality to the theory of belief does not restrict the generality of the theory in any way, as one can choose $\precsim$ to be the full relation in which $\phi \precsim \phi'$ for every $\phi, \phi \in \Phi$. This trivial preference order makes all states equally desirable, and hence makes all states optimal with respect to any set of states.

Simply assuming that beliefs are optimal with respect to a preference order says nothing about from whence the preferences issue. This includes silence on whether the preference order constitutes a fundamental character of the believer that remains the same as beliefs change or instead changes over time. The idea of optimal belief also is silent on what aspects of the form or content of beliefs are distinguished by the preferences. We will consider presently a number of familiar grounds for preferences, and begin with types of preferences that one can regard as expressed within the mental states themselves.

4.3.1 EPISTEMIC PREFERENCES

One can interpret nonmonotonic reasons as expressing preferences over mental configurations in a natural way [24]. In particular, we interpret the reason $A \setminus\setminus B \mid\!\!\!\sim C \setminus\setminus D$ as expressing the preference

$$A\overline{B}C\overline{D} < AB\overline{C}\overline{D} < A\overline{B}C\overline{D} \tag{20}$$

for configurations satisfying $A\overline{B}C\overline{D}$ to configurations satisfying $AB\overline{C}\overline{D}$, *ceteris paribus* [23], and for configurations satisfying either of those conditions to agnostic configurations satisfying $A\overline{B}\overline{C}D$, *ceteris paribus*. That is, if x and x' contain all elements of A, no elements of D, and agree on all elements not in $A \cup B \cup C \cup D$, but x contains all elements of B and none of C while x' contains all elements of C but none of B, then $x \precsim x'$.

Adding optimality with respect to default preferences need not change the conception of belief states, however. In particular, adding optimality with respect to reason preferences to the theory of globally grounded mental configurations described in Section 4.2.3 leaves the sets of extensions invariant, for one can prove that grounded mental configurations are optimal with respect to reason preferences [12].

4.3.2 GROUNDS FOR PREFERENCES

We expect that preferences among beliefs stem from differences in the sources, forms, and content of beliefs, as well as the source and form of the interaction giving rise to the change.

Sources: Preferences based on the sources of information are numerous and familiar. Information influencing belief might come from oneself, through reasoning or perception, or through different modalities within reasoning and perception. Information can also come from other individuals, who stand in various social relationships to the believer, and who act as either authorities in respect to different questions or as observer or informant media through which different types of information flows.

Information influencing belief might also come from groups as well as individuals, most commonly groups to which the believer has a specified relationship. Within this category of sources, one can distinguish information coming from group leaders or non-leader group members; from co-equals, neighbors, friends, family, tribal relations, priests, and religious relations.

Content: Preferences can depend on the type of content or subject matter of the information as well as the source. Information relating to personal, family, and social matters can be treated differ-

ently than information related to sacred, ideological, political, legal, economic, scientific, medical, or fashion. These categories need not exhaust the types of differentiation, either with respect to range of coverage or with respect to more specific subcategories of importance to preferential discriminations.

The psychological content of information might also serve to shape preferences, in that preferences might differ depending on whether the new information comes as a direct change to beliefs or in the form of a change to intentions, desires, or preferences.

Form: Preferences might depend on the form or structure of beliefs as well as the source and content. For example, preferences might depend on the grounding of a belief, with a belief derived strictly from fundamental beliefs preferred to a belief derived strictly from ephemeral beliefs, or derived loosely from a beliefs of mixed character. Within these, preferences might distinguish between conclusions inherent in the logic of belief consistency and closure from those dependent on content-dependent inferences. Preferences might distinguish inferences in one subject area from inferences in another when the abilities of the reasoner differ from one subject to another, perhaps reflecting different dimensions of intelligence. Here one can distinguish the form and content of the grounding from the sources of this form and content.

Preferences might also depend on other aspects of the form of beliefs and mental states, such as whether the belief is an assumption or a derived conclusion, and whether the belief is a component of some identified subtheory or another.

Governmental role: Some preferences will serve to govern conflicts that arise between other preferences. The model here is the occasional conflict between the legal principles of *lex superior* (follow the higher venue) and *lex posterior* (follow the later ruling) in precedence-based legal systems. These two useful principles do not have a definitive outcome when applied to a recent low court ruling that is now seen to conflict with an older high court ruling, so a conflict-breaking metapreference is adopted saying that *lex superior* overrides *lex posterior* in such cases.

Although one might expect that a well-structured believer might employ a system of preferences and metapreferences that always yields a determinate conclusion, achieving such a system of preferences is not at all trivial. It requires substantial manual labor in construction to ensure that no unresolved conflicts result, or some system of automatic conflict resolution that handles all problems in some systematic way. Unfortunately, the latter approach seems to be impossible in general, for the problem of combining multidimensional epistemic preferences of the kind envisioned here produces a version of the group decision problem shown to be unsolvable by Arrow's impossibility theorem [2, 24].

5. The structure of belief change

We divide changes in beliefs into what we will call *motivated* and *accomodative* parts, and differentiate different kinds of such changes according to various qualities, including internality, of rationality, and consciousness.

The motivated part of a change of belief consists of specific properties that are required to hold of the resultant belief state, properties that normally take the form of specific beliefs that are either required to appear or required to not appear in the resultant belief state. For example, in the theory of ideal belief change developed by Alchourrón, Gärdenfors, and Makinson (AGM) [1, 29] considers three types of motivated changes, namely *addition* of some proposition a to the current set of beliefs

even at the cost of making beliefs inconsistent, *contraction* that removes a from the current set of beliefs, and *revision* that adds a while preserving consistency of beliefs, all under the assumption that mental configurations are closed under entailment $x = \mathit{Cn}(x)$, with these operations related by the Levi identity.

Addition:	$x + a = \mathit{Cn}(x \cup \{a\})$
Contraction:	$x \dot{-} a$ removes a from x if possible
Revision:	$x \dot{+} a$ consistently adds a to x
Levi:	$x \dot{+} a \stackrel{\text{def}}{=} (x \dot{-} \neg a) + a.$

In this setting, we regard the proposition a as the motivated change, and identify the accomodative part of the belief change as those changes not stipulated by the motivated part, but which must be made in order to transform the result of enforcing the motivated change into an admissible mental configuration. In the AGM case, beliefs are required to be logically consistent and consequentially closed, so the accomodative change needed when replacing a current belief b is replaced by its contrary $\neg b$ consists of removing other beliefs, such as the belief that $b \wedge b$.

Each of these motivated and accomodative parts of belief change have their own origins and character. We first consider various forms of accomodative change, which we regard as completely internal to the individual or determined by the very constitution of the individual, and then turn to examine the broader forms of motivated change, which we regard as stemming from internal and external sources.

5.1 Accomodative change

The simplest and least problematic form of accomodation corresponds to the addition operation mentioned above, in which one augments beliefs with some new belief and all the consequences flowing from it together with the prior beliefs. In a probabilistic conception of belief, the corresponding operation is Bayesian conditionalization, in which following the stipulation that $Pr(e) = 1$, one replaces the prior probability $Pr(b)$ of each belief b with the conditional probability $Pr(b|e)$.

The applicability of these easy forms of accomodation is limited, however, because, having thought about matters in the past, individuals are likely to have beliefs inconsistent with the new belief. In this case, logical addition produces an inconsistent set of beliefs, and Bayesian conditionalization is undefined, as the required conditionalization by a zero probability proposition would require division by zero.

If one is to not abandon all belief upon discovering an inconsistency of new information with old, one needs to retain some, if not most, current beliefs in changing to account for the new information, and must settle on some principles or means for deciding what beliefs are retained and what beliefs are abandoned.

5.1.1 COHERENT ACCOMODATION

Traditionally belief retention has been viewed in terms of a sort of cognitive inertia or persistence principle in which the grounds for a belief need not be that it follows by some sort of logic or reason, but only that it was believed before, and no reason has been found to abandon it. Simple inertial persistence does not imply any resistance to future abandonment or qualification of the belief, and makes no judgment about which beliefs to abandon when alternative inertial revisions exist. The

problem of determining what beliefs persist and what beliefs need changing has been called the "frame" problem in early studies of automated planning [40, 51, 34].

In the simplest conception of accomodation, one retains all beliefs save those that have been explicitly removed. This is hardly a useful conception of accomodation, as it can leave in place beliefs that singly or jointly imply the removed belief. One thus looks for revisions that move to belief configurations that satisfy the relevant logical and connective coherence criteria.

To characterize coherent belief change with respect to a purely logical conception of coherence, Alchourrón, Gärdenfors, and Makinson (AGM) [1, 29] present the following eight axioms about the notion of contraction, in which the term "theory" refers to a set of beliefs closed under logical entailment.

($\dot{-}1$) $A \dot{-} a$ is a theory whenever A is *(closure)*

($\dot{-}2$) $A \dot{-} a \subseteq A$ *(inclusion)*

($\dot{-}3$) If $a \notin \mathit{Cn}(A)$, then $A \dot{-} a = A$ *(vacuity)*

($\dot{-}4$) If $\nvdash a$, then $a \notin \mathit{Cn}(A \dot{-} a)$ *(success)*

($\dot{-}5$) If $\vdash a \leftrightarrow b$, then $A \dot{-} a = A \dot{-} b$ *(equivalence)*

($\dot{-}6$) $A \subseteq \mathit{Cn}((A \dot{-} a) + a)$ whenever A is a theory *(recovery)*

($\dot{-}7$) $(A \dot{-} a) \cap (A \dot{-} b) \subseteq A \dot{-} (a \wedge b)$ whenever A is a theory

($\dot{-}8$) If $a \notin A \dot{-} (a \wedge b)$, then $A \dot{-} (a \wedge b) \subseteq A \dot{-} a$ whenever A is a theory

These axioms mainly state fairly intuitive conditions, for example, that $A \dot{-} a$ is always included in A ($\dot{-}2$); that contraction leaves A unchanged if $a \notin \mathit{Cn}(A)$ ($\dot{-}3$); that the contraction omits a as long as $\nvdash a$ ($\dot{-}4$); that contractions by equivalent statements yield the same result ($\dot{-}5$); and that adding a to the result of contracting by a only yields conclusions present prior to the contraction ($\dot{-}6$).

Alchourrón, Gärdenfors, and Makinson also present eight axioms paralleling these that characterize the notion of revision, and show that the two sets of axioms are equivalent when connnected by the Levi identity. They also show that these axioms are satisfied by a number of forms of belief revision.

Consider, for example, the set

$$A \downarrow a \stackrel{\text{def}}{=} \max_{\subseteq}(\{B \subseteq A \mid B \nvdash a\}), \tag{21}$$

consisting of the maximal subsets of beliefs A that do not contradict an unwanted belief a, where

$$\max_{\subseteq}(A) \stackrel{\text{def}}{=} \{x \in X \mid \forall y \in X.\ yRx\}. \tag{22}$$

It is easy to see that if A is a theory, so are the elements of $A \downarrow a$. Alchourrón, Gärdenfors, and Makinson give the names maxichoice, partial meet, and full meet contractions, respectively, to contractions that change current beliefs to either one of the sets in $A \downarrow a$, to the intersection of some sets in $A \downarrow a$, or to the intersection of all sets in $A \downarrow a$. They prove that many of the contraction

axioms are satisfied by changing beliefs, and that all are under additional assumptions about the selection of subsets of $A \downarrow a$.

The AGM conception of belief revision treats only definite belief states directly, so one might ask whether the AGM axioms can be recast to apply to indefinite belief states, or even to distributional belief states. AGM do provide one sort of connection between the different conception in their notions of partial and full meet contraction, in which the beliefs following a contraction represent the intersection of all or some or one of the maximal subsets of beliefs consistent with the stipulated contractor.

The question of indefinite belief state belief revision does raise a further possibility not considered by AGM, namely the rationality of an operation one might call *redoing* of prior revisions, in which one switches from one possible revision to another. For example, if the prior contraction settled on one maximal subset of the former beliefs, the redoing operation would switch to some other maximal subset of those same former beliefs. If the prior contraction settled on a partial meet of some of the maximal subsets, the redoing operation might change to the partial meet of some selection of the maximal subsets. Moreover, a hesitant believer might perform a sequence of redoings, possibly ending up with the same beliefs as after the initial contraction. Outcomes of the redoing operation, unlike those of addition, contraction, and revision, clearly depend on the history of changes to beliefs in a nontrivial way. Although one can formalize the redoing operation in the AGM framework, the motivations for doing so are greater when one considers foundational approaches to belief revision.

More generally, as one expands the conception of a belief state from a set of beliefs or statements of the same character to a structure in which some beliefs are base beliefs, some are conclusions, some are assumptions, some represent contradictions, and some represent reasons, one opens the door to a wider range of stipulations. One can then distinguish adding a base belief from adding a conclusion, and distinguish defeating a specific assumption or inference from simply contracting by some unwanted conclusion.

5.1.2 PREFERENTIAL ACCOMODATION

The AGM characterization of belief revision aims to be more inclusive than Rescher's [50] notion preferred maximal alternatives. One can connect the two perspectives by considering again the notion of preference orders over belief states and the notion of rational choice among available contractions. To do this, we define the operation $\hat{-}$ of maximally preferred contraction by

$$A \mathbin{\hat{-}'} a \stackrel{\text{def}}{=} \max_{\prec}\{Cn(B) \mid B \subseteq A \wedge B \nvdash a\} \tag{23}$$

$$A \mathbin{\hat{-}} a \stackrel{\text{def}}{=} \begin{cases} \{A\} & \text{if } \vdash a \\ \{A\} & \text{if } A \mathbin{\hat{-}'} a = \emptyset \\ A \mathbin{\hat{-}'} a & \text{otherwise.} \end{cases} \tag{24}$$

The result of rational contraction thus represents the maximally preferred subsets of beliefs out of all consistent subsets that avoid the stipulated contractor. Doyle [17, 16] proves that this *de minimis* notion of rational contraction satisfies some but not all of the AGM postulates. The difference arises in part because the pure theory of rational revision allows smaller sets of beliefs to be preferred to larger sets. Formally, we say that a preference order $\precsim$ is *informationally monotone* iff $A \precsim B$ whenever $Cn(A) \subseteq Cn(B)$, that is, $A \precsim B$ whenever $B \vdash A$. One can then prove that rational

choice contraction with respect to an informationally monotone preference order satisfies all the AGM axioms [17, 16].

The notion of informational monotonicity of preferences over belief states has at least three natural interpretations. The first interpretation embodies the epistemological principle that knowledge is good, and that more knowledge is better. To the extent that belief is a proxy for knowledge, this principle implies that more beliefs are better than fewer beliefs, ignoring differences of content. The second interpretation restates the first from a utilitarian or economic point of view, embodying the economic principle that belief is a valuable resource, and that therefore beliefs that represent the product of past work should not be discarded needlessly. The third interpretation, which we return to later, regards beliefs as subject to a principle of inertia, according to which beliefs are maintained as long as they were believed earlier and not removed. Although the epistemic and utilitarian interpretations of informational monotonicity are both based on a notion of the good, the inertial interpretation clearly represents a principle of a different character that makes no reference to any conception of the good.

One obtains a somewhat more complicated conception of belief change when one considers nonlogical connections and the preferences they express. Nonlogical connections violate the monotonicity requirements of the AGM revision principles, as do their associated epistemic preferences. In this case, and in all cases in which preferences can change with mental configuration, one must decide which preferences shape preferential revision, whether the preferences held prior to the change, or the preferences held once the change has been effected. For changes judged by prior preferability, one chooses B that maximizes prior preferences $\precsim_A$ seen in A, that is, for each $C \in A \downarrow a$, $C \precsim_A B$. For changes judged by posterior preferability, which Jeffrey [35] terms "ratified" choices, one chooses B that maximizes posterior preferences $\precsim_B$ seen in B, that is, for each $C \in A \downarrow a$, $C \precsim_B B$.

5.1.3 CONSERVATIVE ACCOMODATION

One obtains a still more complicated conception of belief change when one considers grounded mental configurations and looks for changes in the set of grounded extensions of the base beliefs.

Base beliefs appear as unremovable as tautologies when specifying revisions in terms of changes to conclusions, while revisions mediated via changes to base beliefs underdetermine changes to conclusions in the sense that removing some base belief in order to remove some conclusion it supports may fail, leaving the unwanted conclusion still believed through support tracing to some unremoved base beliefs.

In the setting of changes to base beliefs, one looks to describe changes to conclusions in terms of smoothest path or minimal change principles.

For definite configuration states, when changing base beliefs to m' while in one extension $x \setminus\setminus y \in \alpha(m)$, one must change conclusions to some $x' \setminus\setminus y' \in \alpha(m')$. If one allows this choice to be made without reference to the preceding conclusions, one can make arbitrarily large changes in beliefs, even if $x \setminus\setminus y \in \alpha(m')$. The simplest sort of change principle is to minimize set differences in changes to the *In* and *Out* components of the extension, that is, choose $x' \setminus\setminus y' \in \alpha(m')$ such that for each $x'' \setminus\setminus y'' \in \alpha(m')$, one has $x + x' \subseteq x + x'' \rightarrow x' = x''$ and $y + y' \subseteq y + y'' \rightarrow y' = y''$.

For indefinite configuration states, one simply adds together the possible minimal change configurations available from each of the preceding configurations. If we write $\mu(m, s, x, y)$ to denote the set of configurations that constitute minimal changes from (m, s, x, y), then we can extend μ to

a function on states. For definite configuration states, states are isomorphic to configurations, so

$$\mu(\psi) = \mu(\beta(\psi)). \tag{25}$$

For indefinite configuration states, the additivity condition just proposed can be expressed formally as

$$\mu(\psi) = \bigcup_{\phi \in \beta(\psi)} \mu(\phi). \tag{26}$$

Of course, minimality of changes can be viewed in other ways as well, such as maximizing preferability of the change. There is no necessity to regard minimal change measures as identical to preferences, however, so one can consider composite belief change approaches in which one seeks minimal changes among the preferred extensions, or seeks preferred extensions among the minimal changes.

5.2 Motivated change

Where accomodative change concerns how to update beliefs once it is decided that an update must be performed, motivated change concerns whether to update beliefs and what sort of update to perform.

The simplest sorts of motivated changes concern whether or not to accept new information provided by some source internal or external to the individual. In this case, the response to new information is not mere conditionalization, but first of all a decision about whether to accept the new information in full or in part, or whether to reject it entirely. It is secondarily a decision about how to accommodate the portion to be accepted, about how to change beliefs in light of the acceptance decision.

Human society provides important examples of external motives for belief change. Social psychologists, for instance, have studied the influence of peer pressure, in which the agreement of one's friends on some question or taste influence one to change so as to agree with them. Wicker [60] has developed a formalization covering a variety of such mechanisms of social influence, in which multiple external groups and individuals influence an individual through multiple different mechanisms. We will not examine such mechanisms in detail, but instead will examine such influences by regarding them as forces acting on the individual. In this setting, each different influence mechanism generates a different type of motive for change on the individual or group. Other external motives include information provided by perception.

Internal motives for change include deliberate decisions of the individual to adopt or abandon some belief, wanton volition to satisfy desires, the action of mental habits, and the results of reflection and other self-observations.

Individuals of different constitution can generate different internal motivations for change, and can respond differently to the same external or internal motivations. For example, the "organization man" might conform his beliefs to those common among his co-workers, and the eccentric might choose divergent beliefs to escape the suffocation of conformity. As with external motives, internal motives for change can come through multiple mechanisms, some of which involve inference, some habits, and some goal-seeking or intention-following.

Both acceptance and accomodation decisions depend on many factors, These decisions depend on the type of belief to be accepted, on the specific belief in question, on other beliefs, and on the beliefs of others. Indeed, these decisions can depend on the alternatives available, in that the

decision to accept can be shaped by whether accomodating the information is even possible for the individual or group.

For present purposes, we consider only motives for change that can be summarized in terms of addition or removal of specific beliefs or mental elements to conclusions, base beliefs, or other parts of mental configurations. Such changes include those due to the individual observing that some fact b about the world or about sensations being experienced is true (or false), or due to being told by some other individual or group that b is true, or being told a reason a reason for believing that b is true, or by observing others act as though b is true or as though they believe b is true.

5.3 Rational change

The notions of motivated and accomodative change do not presume any notion of rationality of decision or accomodation, although rationality is the focus of many studies of belief change.

Rationality of internally-motivated change is reasonably viewed in classic decision-theoretic terms as consisting of selection of a (possibly empty) set of beliefs as the target for change in which the selection is of maximal desirability or expected utility to the individual, either with respect to the beliefs and preferences holding before the change (prior rationality) or with respect to the beliefs and preferences holding after the change (posterior or ratified rationality [35]). The question of rationality of externally-motivated change does not arise except possibly with respect to decisions made by external individuals that lead to the imposition of the external changes.

Rationality of accomodation can be viewed in the same way as rationality of internally-motivated change, as choosing the accomodation of maximal desirability from among those available. This conception, however, undercuts the distinction between motivated and accomodative changes, for if one can choose both motivated targets and their accompanying accomodations via the same type of rational decision, why not simply choose an overall change rationally and be done with it? Indeed, the primary reasons for distinguishing motivated and accomodative changes has to do with limitations on the self-knowledge and reasoning powers of the individual. For individuals of realistic mental capability, it is reasonable to regard them as making rational decisions about what assumptions, subgoals, and subplans to adopt in pursuing desires or carrying out plans, but it is not reasonable to regard them as being able to forsee all the consequences of these motivated changes. One thus looks for conceptions of accomodative change that do not involve comparison of the overall desirability of complete changes of mental state.

The conceptions of rational accomodation examined in the following thus judge rationality by potentially weaker standards than decision-theoretic rationality with respect to all the beliefs and preferences of the individual. The most prominent conceptions involve standards reflecting conservation or minimal change principles, variously conceived in informational, economic, and mechanical terms, that is, as minimizing loss of beliefs, loss of cognitive capital, change in mental inertia or energy, or as maximization of desirability with respect to limited types of preferences.

In the setting of rational motivations and minimal-change accomodations, one can recast the division of change into motivated and accomodative parts as the division of reasoning into *progressive* and *conservative* reasoning [13, 15], usually but not always representing the distinction between intended changes and unintended consequential changes.

One might also differentiate motivated and accomodative changes according to degree of consciousness of the change, for example, regarding changes motivated by deliberate decisions as conscious changes, and regarding loss-minimizing conservative accomodations as unconscious

changes, but we do not attempt to treat questions of consciousness or unconsciousness of change here.

5.4 Mechanical change

The notions of motivated and accomodative changes discussed in the preceding concern global changes from one belief configuration to another, with no discussion of exactly how such changes take place or of how long they take. To discuss the fine structure of processes of belief change over time, we employ the concepts of mechanics [21] to recast elements of the preceding discussion.

5.4.1 MASS, POSITION, MECHANICAL CONFIGURATION

We obtain mechanical notions of mass and position from the grounded mental configurations $(m, s, x, y) \in \mathbb{D} \times \mathbb{S} \times \mathbb{D}^2$ presented earlier.

We interpret the last three components $\chi = (s, x, y)$ as the *position* of the individual. This usage fortuitously accords well with one sense of the word "position" in ordinary speech, as one's beliefs regarding some issue are often called one's position on the issue.

We interpret the m component of mental configurations as the *mass* of the individual. Mental mass constitutes that portion of memory that persists independently of motion. In psychological terms, belief mass forms part of what we think of as memory, along with some sorts of configurational information. Mass, like configurations of a rigid body, persists through inertial motion, but mass, unlike configurations, cannot change merely through motion. Just as different individuals can inhabit different types of belief states, one can consider them as exhibiting mass in different ways, including masses that fall in discrete or continuous spectra. When considering beliefs represented in the space $\mathbb{D}$, we regard discrete mass values as represented by the same space, namely $\mathbb{D}$. For continous mass values, we will normally consider values in $\mathbb{R}^+ = \{r \in \mathbb{R} \mid r \geq 0\}$, or sometimes values in $[0, 1]$. The discrete position spaces considered here call for a different dimension of mass for each dimension of the space of positions, unlike the case in ordinary physics in which a body has the same mass in all dimensions.

We obtain the notion of mechanical configuration by extending mental configurations with a third component $\dot{\chi} = (\dot{s}, \dot{x}, \dot{y}) \in \mathbb{S} \times \mathbb{D}^2$ that represents a velocity value. Comparing mental and mechanical configurations, we see that mental configurations are synchronic, while mechanical configurations are diachronic.

We regard the pair $p = (m, \dot{\chi})$ as a momentum value. Because mass values are not scalars of the vector space of positions in the discrete position spaces considered here, momentum calls for a different dimension of mass for each dimension of the space of positions, unlike the case in ordinary physics in which a body has the same mass in all dimensions.

With this identification, we thus can write a mechanical configuration in two forms, as $(m, \chi, \dot{\chi})$ or as (χ, p). The latter expression corresponds to the familiar notion of dynamical state in Hamiltonian mechanics. We will write Φ in the following to denote the set of all mechanical configurations $(m, \chi, \dot{\chi})$ in $\mathbb{D} \times (\mathbb{S} \times \mathbb{D}^2)^2$. We write $\mathbf{0}_\chi$ to denote the zero vector in $\mathbb{S} \times \mathbb{D}^2$.

Recalling our earlier discussion of non-grounded configurations, we see that $m = \mathbf{0}$ and $s = \mathbf{0}$ in pure coherence configurations. For terminology perhaps more in line with common speech, one might call the combination (m, χ) of mass and position a pure state of mind, and the combination $(m, \dot{\chi}, \chi)$ of position and momentum a pure state of reasoning.

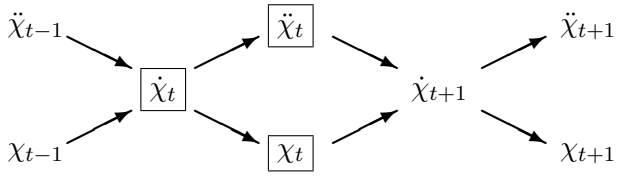

Figure 7: The kinematical relationships among position variables in time. The boxed quantities denote conventional labels for the quantities of interest at instant t, with a reasoning agent observing χ_t and $\dot{\chi}_t$ and choosing $\ddot{\chi}_t$.

5.4.2 MECHANICAL MOTION AND STATES

We write $\mathcal{T}$ to denote a linearly ordered set of times or temporal indices for use in describing changes of belief and mental state over time. We write $[t, t']$ to denote the closed interval of time starting at t and ending at t', and write (t, t') to denote the open interval of times occuring after t and before t'. When t is an element of a discrete sequence of times, we write $t - 1$ and $t + 1$ to denote the immediately preceding and succeeding instants.

We define a *(lineal)* history or trajectory of an individual over a temporal interval I to be a function $h : I \to \Phi$, that is, a function $h : I \to \mathbb{D} \times (\mathbb{S} \times \mathbb{D}^2)^2$. When each history in a set H of lineal histories is defined over the same temporal interval, that is, when $I_h = I_H$ for each $h \in H$, we say that H constitutes a *nondeterministic history*. Each nondeterministic history H over I_H induces, for each discrete sequence of instants $I_D \subseteq I_H$, a discrete correspondence (set-valued function) $\Delta_H : \Phi_i \times I_D \to \mathcal{P}(\Phi)$ over configurations and instants defined so that at every instant t in the sequence I_D, we have $\phi' \in \Delta_H(\phi, t)$ iff there exists some $h \in H$ in which $h(t) = \phi$ and $h(t + 1) = \phi'$.

As in the relations of definite, indefinite, and distributional states of individuals to mental configurations, we regard mechanical states of individuals as taking definite, indefinite, and distributional forms. Each nondeterministic history, therefore, constitutes a lineal histories of indefinite states.

We write χ_t to denote the indicated position at discrete instant t, and $\dot{\chi}_t$ to denote the indicated trailing velocity at that instant. For χ to constitute a mechanical history, the indicated velocity value $\dot{\chi}_t$ must match the actual velocity determined by χ, that is it must be the case that $\dot{\chi}_t = \chi_t - \chi_{t-1}$. This trailing velocity corresponds to the change signals used in some automated reasoners in triggering rules [9]. The leading acceleration $\ddot{\chi}_t = \dot{\chi}_{t+1} - \dot{\chi}_t$ reflects the additions and removals indicated by steps of reasoning. We depict these kinematical quantities in Figure 7. We write m_t to denote the mass at discrete instant t, and $\dot{m}_t = m_{t+1} - m_t$ to denote the leading mass flux or change of mass. Similarly, we write p_t to denote the momentum at t, and $\dot{m}_t = m_t - m_{t-1}$ to denote the momentum flux or change of momentum. By the above identifications, a momentum flux value $\dot{p} = (\dot{m}, \ddot{\chi})$ combines a mass flux value with an acceleration value.

We say that an indefinite state has definite mass, position, or velocity, respectively, if all mechanical configurations in the state have the same mass, position, or velocity. We say that a distributional state has definite mass, position, or velocity, respectively, if it assigns probability 1 to some mass, position, or velocity. In the following we will restrict attention to states exhibiting a definite mass, but states without a definite mass might be of interest in understanding the possible foundations for

an observed position, or in obtaining a mechanical understanding of Dempster-Shafer [54] evidential mass.

For indefinite configuration states, one simply adds together the possible minimal change configurations available from each of the preceding configurations. If we write $\mu(m, s, x, y)$ to denote the set of configurations that constitute minimal changes from (m, s, x, y), then we can extend μ to a function on states. For definite configuration states, states are isomorphic to configurations, so

$$\mu(\psi) = \mu(\beta(\psi)). \tag{27}$$

For indefinite configuration states, the additivity condition just proposed can be expressed formally as

$$\mu(\psi) = \bigcup_{\phi \in \beta(\psi)} \mu(\phi). \tag{28}$$

Note that the set of velocities exhibited in this superposition state only includes the changes corresponding to indeterministic transitions from definite configurations, and need not include the differences of all positions exhibited in the state.

5.4.3 FORCE

We denote the force at an instant by f_t. In mechanics, force values have the same dimensionality as momentum values, in order to satisfy the Euler-Newton law

$$f = \dot{p}, \tag{29}$$

so in the present setting, therefore, we identify the set of possible force values to be $\mathbb{D} \times (\mathbb{S} \times \mathbb{D}^2)$ and have

$$f_t = \dot{p}_t = (\dot{m}_t, \ddot{\chi}_t). \tag{30}$$

Of course the dimensionality of forces varies with the conception of the nature of individuals and their states.

We describe influences of one set of individuals on another in terms interactions or forces that apply concurrently. Drawing on the formulation by [43], we say that a system of forces f is a function that maps each pair g, g' of separate groups ($g \cap g' = \emptyset$) to a force value $f(g, g')$, interpreted as the force exerted by g' on g, such that

$$f(g_1 \sqcup g_2, g) = f(g_1, g) + f(g_2, g) \tag{31}$$

$$f(g, g_1 \sqcup g_2) = f(g, g_1) + f(g, g_2) \tag{32}$$

We write $f(i, g)$ as shorthand for the force $f(\{i\}, g)$, meaning the force exerted by group g on individual i. Similarly, we interpret $f(i, i')$ as shorthand for $f(\{i\}, \{i'\})$. Mechanics typically assumes additional properties of systems of forces beyond the additivity required in the preceding. In particular, force systems are normally required to be balanced, in the sense that

$$f(g, g') = -f(g', g) \tag{33}$$

for all pairs of groups, and the complete system of forces is assumed to be the sum

$$f = f^{\mathrm{B}} + f^{\mathrm{C}} \tag{34}$$

of systems of body and contact forces f^{B} and f^{C}. For continuum bodies, mechanics extends the additivity requirements by assuming that forces are integrals of force densities that give rise to the traction (contact force density) and the stress tensor.

The additivity of force systems lets us divide the total force

$$f_t = f_t^{\mathbf{a}} + f_t^{\mathbf{s}} \tag{35}$$

on the individual into the *applied* force $f_t^{\mathbf{a}} = f_t(i, i^{\mathrm{e}})$ of environment on the individual with the *self* force $f_t^{\mathbf{s}} = f_t(i, i)$ of the individual on itself. The simplest kind of self-generated force is the inertial force $-\dot{p}$ generated by mass. The Euler-Newton equation $f = \dot{p}$ equates the total noninertial force f on a body with the negative of its inertial force $-\dot{p}$, defined as its change of momentum in an inertial frame of reference. The total force acting on a body is the sum of the applied forces and the inertial forces, and hence is zero.

Only a force with a mass-flux component can change the mass of an individual, meaning no mere change of position can remove or augment mass. We thus distinguish two classes of forces, *mass* forces $f = (\dot{m}, \mathbf{0})$ involving only a change of mass, and *spatial* forces $f = (\mathbf{0}, \ddot{\chi})$ involving no change of mass. In considering changes to grounded configurations, mass forces correspond to stipulating a change in the base beliefs, and spatial forces correspond to stipulating changes in conclusions.

Different interpretations apply to different components of $\dot{p} = (\dot{m}, \ddot{\chi})$. The mass flux component $\dot{m}$, as an element of $\mathbb{D}$, represents both a set $m\dot{m}$ of additions and a set $\overline{m}\dot{m}$ deletions from the current mass value m. Starting from sets m^+ and m^- of beliefs that should and should not be present in the new mass, one obtains the sets of additions $\dot{m}^+ = \overline{m}m^+$ and deletions $\dot{m}^- = mm^-$ and adds them together to get the mass flux $\dot{m} = \dot{m}^+ + \dot{m}^-$.

The acceleration component $\ddot{\chi} = (\ddot{s}, \ddot{x}, \ddot{y})$ provides only one means for specifying changes in conclusions, for given the current mechanical state $(m_t, \chi_t, \dot{\chi}_t)$ one can derive $\ddot{\chi}_t$ either from a stipulated value for χ_{t+1} or from a stipulated value for $\dot{\chi}_{t+1}$ using the kinematic relations between these values. For example, from a stipulated value χ_{t+1}, we obtain $\dot{\chi}_{t+1} = \chi_{t+1} - \chi_t$ and $\ddot{\chi}_t = \dot{\chi}_{t+1} - \dot{\chi}_t$.

Standard practice in belief revision and in reasoning systems is to specify changes in terms of belief stipulated to hold in either the base or conclusion sets. We have seen how to translate such stipulations of base beliefs into mass flux components of forces, but there is no simple translation of stipulated conclusions into positions (s_t, x_t, y_t) and velocities $(\dot{s}_t, \dot{x}_t, \dot{y}_t)$. Stipulating that the new conclusions x_{t+1} should contain particular beliefs and that the new non-conclusions y_{t+1} should contain other particular beliefs does not determine these sets uniquely, much less specify any changes to s_t. The purpose of RMS and related revision systems is to compute full changes to conclusions and supports from the partial specification given by stipulated beliefs and nonbeliefs. These changes then determine the full acceleration values according to the kinematic definitions.

5.4.4 ELASTIC ACCOMODATION AND REASON FORCES

Although one can motivate the conservative principle underlying coherent accomodation in terms of inertia, a passive resistance to any change, the role played by logical and psychological connections between beliefs in coherent and grounded accommodation suggest more active forms of resistance, in which each belief held is motivated and maintained by present reasoning, and in which these dependencies resist change, in that effecting some change means overcoming alternate derivations

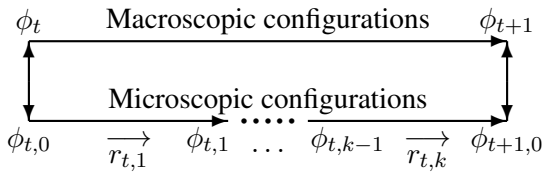

Figure 8: Reasoned decomposition of a transition between macroscopic configurations ϕ_t and ϕ_{t+1} into a series of "microtransitions" between microscopic configurations $\phi_{t,i}$ effected by a series of reason applications $r_{t,i}$.

as well as the current support. One can view active resistance of this sort as a form of elasticity of the mental material.

From the point of view of an external observer, the RMS moves from one complete globally grounded belief configuration to another. We call these externally visible configurations *macroscopic* configurations, and changes between them macroscopic changes. These external appearances hide the fine structure of an externally invisible unlabeling and relabeling process by which one coherent grounded configuration is transformed into another by a sequence of *microscopic* transitions between incomplete microscopic configurations. In the following, we indicate how one can think of macroscopic configurations as "relaxed" or equilibrium configurations, with stable belief ascriptions, and microscopic configurations as nonequilibrium configurationss passed through on a mechanical path to an equilibrium configuration, as depicted in Figure 8.

In the classical RMS revision model, revision starts with a stipulated change to the base beliefs and a complete mental configuration in which each possible belief is labeled either *In* or *Out* and a supporting reason is identified for each *In* belief. If the reason to be removed from the base set is invalid in the current configuration, it is merely added or removed from the base beliefs and the revision is done. If the reason is to be added and is valid and supports some conclusion that is currently *Out* but which does not appear in the defeaters of the supporting reason for some other belief, then the conclusion is added to the set of *In* beliefs and the support assignment is augmented to indicate the new reason as the support for this new conclusion. If some existing conclusion had support which did have the new conclusion as a defeater, then an extended relabeling process is initiated, in which the RMS temporarily unlabels conclusions with supports derived from the changed belief and reexamines reasons to see if alternative supports can be found. At the end of the process, the RMS either has a complete globally grounded configuration or indicates that no such configuration exists.

This original approach of propagative unlabeling and relabeling is not the only means by which one might effect reasoned revisions. A very different procedure, perhaps more natural in the setting of mechanics, is one of concurrent relabeling, which also propagates relabeling, but in which any frontier node can start labeling, yielding a set of overlapping relabeling processes which, like ripples from several pebbles tossed in a pond, expanding, superimpose, and reflect off barriers.

We do not treat all details of reasoned revision here, in part because different methods for performing such revisions correspond to different types of forces, and in part because the most familiar methods involve considerable complexity of detail. We instead merely give examples of particular types of forces generated by reasons at different points of the revision process, expressing these

forces in terms of a mechanical state $(m_t, (s_t, x_t, y_t), (\dot{s}_t, \dot{x}_t, \dot{y}_t))$. We do not describe here how these forces are generated at the appropriate times.

In the RMS conception, the RMS is a subsystem of a larger reasoning system. In the terms used here, the RMS is a subbody or member of the larger individual. The larger individual interacts with the RMS by exerting forces on it to change base reasons and to query the status of different possible beliefs. We do not treat queries here.

We regard addition or removal of a base reason as mediated through a mass force

$$f_t^{\mathbf{a}} = (\dot{m}_t, \mathbf{0}_\chi) \tag{36}$$

that changes the set of base reasons from m_t to m_{t+1}.

We regard the forces of reasoning and some forms of learning as spatial self forces. For present purposes, we assume that these self-forces can depend on the mechanical state, that is,

$$f^{\mathbf{s}}(m_t, (s_t, x_t, y_t), (\dot{s}_t, \dot{x}_t, \dot{y}_t)) = (\mathbf{0}, (\ddot{s}_t, \ddot{x}_t, \ddot{y}_t)). \tag{37}$$

In the following, we regarding $f^{\mathbf{s}}$ as a sum $f_1^{\mathbf{s}} + \cdots + f_n^{\mathbf{s}}$ of several contributing self-forces, namely ones connected with addition of beliefs and the application of reasons in the course of propagative labeling and unlabeling.

We regard the forces mediating propagation of unlabeling and relabeling as spatial self forces, with each reason r generating a force

$$f_r^{\mathbf{s}}(m_t, (s_t, x_t, y_t), (\dot{s}_t, \dot{x}_t, \dot{y}_t)) = (\mathbf{0}, (\ddot{s}_t, \ddot{x}_t, \ddot{y}_t)) \tag{38}$$

with values as follows. To simplify the exposition, we treat only $\ddot{x}_t$ and $\ddot{y}_t$ in the following, and do not give details for $\ddot{s}_t$.

If $r = A \setminus\setminus B \mid\!\!\vdash C \setminus\setminus D$ is invalid in the configuration (m_t, s_t, x_t, y_t), that is, if $A_r\overline{x}_t + B_r\overline{y}_t \neq \mathbf{0}$, then the force of labeling or unlabeling is the null force $(\mathbf{0}, \mathbf{0}_\chi)$. If r is valid, we have the following forces generated by addition and removal, where addition and removal are indicated by the presence of r in $\ddot{x}_t$ or $\ddot{y}_t$.

If the reason is being added, we have

$$\ddot{x}_t = C_r\overline{x}_t - D_r x_t - \dot{x}_t \tag{39}$$

$$\ddot{y}_t = D_r x_t - C_r\overline{y}_t - \dot{y}_t, \tag{40}$$

while if the reason is being removed, we have, for $s_t^{-1}(r) = \{e \in \mathcal{D} \mid s_t(e) = r\}$,

$$\ddot{x}_t = s_t^{-1}(r)C_r\overline{x}_t - s_t^{-1}(r)D_r x_t - \dot{x}_t \tag{41}$$

$$\ddot{y}_t = s_t^{-1}(r)D_r x_t - s_t^{-1}(r)C_r\overline{y}_t - \dot{y}_t. \tag{42}$$

5.4.5 INERTIAL CHANGE

We can identify another form of belief change beyond motivated and accommodative belief change with the aid of mechanics, namely the notion of nonstationary inertial belief change. Inertial motion consists of constant linear motion, constant rotation, or the combination of the two, in which beliefs change without the application of any force. We will look at nonstationary linear inertial motion in Section 6.2, in which the net effect of the linear motion is to introduce uncertainty about the most

recently changed beliefs. More interesting forms of inertial motion involve rotation. As with linear inertial motion, inertial rotation also introduces uncertainty. A natrual form of this uncertainty is cycling through alternative assumptions, with the overall rotation the product of smaller rotations about multiple axes corresponding to different decisions. We will not treat this in detail here. See [21] for some additional discussion.

6. The difficulty of belief change

With this variety of forms of belief change, we turn to means for assessing the difficulty of changing beliefs. We first consider qualitative comparisons of degree of difficulty based on the notions of epistemic entrenchment, and then turn to quantitative comparisons based on the notion of mechanical work.

6.1 Qualitative comparison via entrenchment

Gärdenfors and Makinson [31] translate the AGM structure of belief revision into an ordering of beliefs according to *epistemic entrenchment*, which orders beliefs by relative ease of abandonment. A belief a is less entrenched than b, written $a < b$, if one gives up a before b when one cannot hold both, defined formally by

- $a \leq b$ if $a \vdash b$, and
- $a \leq b$ iff $a \notin A \dot{-} (a \wedge b)$ or $\vdash a \wedge b$.

Gärdenfors and Makinson [31] proposed the following axioms for entrenchment and show their equivalence to the AGM axioms for contraction.

($\leq$1) If $a \leq b$ and $b \leq z$, then $a \leq z$

($\leq$2) If $a \vdash b$, then $a \leq b$

($\leq$3) Either $a \leq a \wedge b$ or $b \leq a \wedge b$

($\leq$4) If A is consistent, then $a \leq b$ for all b iff $a \notin A$

($\leq$5) If $a \leq b$ for all a, then $\vdash b$

Although the axioms characterizing ideal epistemic entrenchment yield the same conception of revision as do the AGM revision axioms, the axioms make no explicit separation between types or sources of beliefs, but instead only presume differing levels of entrenchment. The entrenchment order itself is left formally exogenous to beliefs, desires, intentions, and other mental attitudes. Thus the axioms alone say nothing about the origin of entrenchment, about why one order and not another, or about whether or how entrenchment changes over time or with beliefs. Because of this silence, one must look elsewhere for the grounds of entrenchment. We now consider origins for entrenchment in dependencies among beliefs, motives for belief, and preferences among beliefs.

6.1.1 ENTRENCHMENT FROM FOUNDATIONS AND PREFERENCES

Gärdenfors [30] proposed that one can interpret coherence theories of belief revision so as to identify reasons for belief, the simplest connection being to say that b is the (only) reason for c if both are believed and $b \vee c \leq b$, so that c is removed whenever b is. Doyle [18] took issue with Gärdenfors' claims, and argued that reasons form the practical expression of entrenchment order.

The structure of globally grounded extensions provides several means for analyzing properties of the resulting entrenchment order.

The first structural analysis of grounded extensions is to stratify them by inference level [11, 19]. In this stratification, the base beliefs appear in level 0. All beliefs derived by application of a single reason from level 0 beliefs, such as unconditional assumptions, appear at level 1, and beliefs supported only by reasons with antecedents appearing first at level n appear at level $n+1$. In general, a belief has level n just in case the minimal height of the belief in any grounded configuration is n, that is, a chain of at least n reasons is needed to derive the belief in any grounded configuration.

Information about the level of a belief provides some information about entrenchment, but not a complete characterization. Because all base beliefs in m appear at level 0, levels provide no comparisons of differences in entrenchment among base beliefs. The same holds true for condition-free assumptions, which appear (along with other conclusions) at level 1. On the other hand, assumptions and other beliefs can appear at any level of the hierarchy, with higher-level assumptions conditioned on lower-level assumptions and beliefs.

Deriving entrenchment from foundations thus helps answer one of the questions left untreated in the AGM framework, namely where entrenchment comes from and how it changes. Ordinary reasoning changes the foundations of belief, either by adding new reasons from simple inferential processes or from chunking and related learning mechanisms. Reason-base restructuring of the RMS variety also changes foundations and hence entrenchment.

Something like these differences in level are exploited in the RMS dependency-directed backtracking procedure, which effects reasoned contraction by recursively tracing the derivations underlying the belief targeted for removal. Reasoned contraction regards the first assumptions reached in this inverse-derivation traversal of the derivations as the least consequential, and chooses to defeat one of these "uppermost" assumptions as a way of securing a change-minimizing revision. In this process, the entrenchment level of a belief is inversely related to its height in the assumption graph. The RMS conception of entrenchment is not the same as the AGM notion, however, in that the

	Base reasons	Conclusions
	$\Vdash a$	a
+	$\Vdash b$	a, b
+	$a, b \Vdash c$	a, b, c
-	$\Vdash a$	b
+	$\Vdash a$	a, b, c
-	$\Vdash b$	a
+	$\Vdash b$	a, b, c
-	$\Vdash a$	b
-	$\Vdash b$	

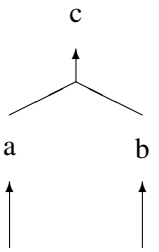

Figure 9: Indirect contraction

RMS conception concerns only derivations in a single grounded configuration, so that an assumption that appears at one level in one grounded configuration might appear at a different level in a different grounded configuration with no change in base beliefs. The full level analysis developed in [11, 19] corresponds more closely to the AGM conception by assigning levels based on the first level at which a belief appears in any grounded extension. Even that notion, however, does not fully match that of AGM entrenchment, in that two beliefs can be both appear at the same level yet be completely independent of each other.

One might also consider incomplete stratifications. For example, in locally grounded configurations, stratification from base beliefs need never reach beliefs appearing in self-supporting cycles, so beliefs in these cycles or dependent on them would not be assigned a level. One might extend these levels to more complete ones by considering relative levels, in which beliefs in one cycle rank less than beliefs in another cycle if some belief in the other cycle appears in a stratification that adds beliefs in the first cycle as assumption.

Base beliefs and unconditioned assumptions represent just the two most obvious cases in which differences in entrenchment require preferences among beliefs. McAllester [39] developed the idea of a preference ordering of base beliefs or assumptions, but this suffices only if all reasons represent monotonic logical inferences. Without this restriction, it is useful to employ preferences ordering assumptions at different levels, such as those expressed in general nonmonotonic reasons.

del Val [8] pursued these questions, and showed the equivalence of coherence and foundations under the following conditions: that reasons are beliefs, that is, there are no nonmonotonic rules; that reasons generate conclusions by deductive closure alone, that is, only standard logic applies; that logically equivalent conclusions have identical bases; and that there is an entrenchment ordering on foundational beliefs. Put another way, the foundations approach is the same as the AGM conception if by foundations is meant simple deductive closure of ordered sets of axioms. This simplified deductive picture is close to the view of revision proposed by McAllester [39] in his TMS, an approach later called LTMS by Forbus and de Kleer [28].

The del Val result provides justification for the claims of both Gärdenfors and Doyle, but does so by assuming away most of the important characteristics of realistic forms of belief revision. Realistic belief revision involves finite and nondeductive reasoning in essential ways, corresponding to finite structure in neuronal embodiments, and to finite representations in computational embodiments. The del Val assumptions thus do not fit realistic minds. One finds more realistic relations between foundations and coherence theories by reconstructing the concept of epistemic entrenchment in the setting of finitary and nonmonotonic foundational structures.

The role of preferences in reasoning provides further information for understanding how entrenchment arises and changes. Reasoning that involves constructing new nonmonotonic reasons implicitly changes the preferences of the individual, and hence entrenchment comparisons. More generally, one can regard problem solving goals as expressing preference information as well [59], and thus regard the reasoning involved in subgoaling problem decomposition as changing goals and possibly entrenchment relations. In particular, assumptions related to the tentative solutions being explored will have a different status than assumptions unrelated to the current focus of problem solving.

The reframing of theories common in mathematics provides an example of problem-specific changes in entrenchment. In this process, mathematicians prove theorems about some subject, and later come to realize that the concepts developed in the theorems actually have broader application and interest in their own right. The mathematicians then interchange the axioms and theorems.

Some of these derived concepts are introduced as new foundational concepts, with the theorems now used as axioms to characterize them, and the former axioms are now proven as consequences of the new axioms. To take a trivial example, one can base the definition of a triangle either on axioms about having three angles or on the basis of axioms about having three sides, and prove each of these definitions as a theorem from the other. A less trivial mathematical example is that after proving separation properties of sets of real numbers and functions, mathematicians made these separation properties into the axioms of topology. Or for an example from economics, one can define rational actions as those maximizing expected utility, based on axioms for preferences and probabilities, or can turn around and use the facts of decisions of assumed rationality as the axioms and derive preferences and probabilities. In each of these, the notion of foundation of belief depends on purpose of reasoning, so in general the notion of entrenchment is relative to how one chooses to represent the problem at hand.

6.1.2 ENTRENCHMENT MODALITIES

Extending the notion of entrenchment to belief states constructed from base beliefs via limited logics and nonmonotonic interval conditionals poses problems right from the start in that the result of contraction need not be a subset of the uncontracted beliefs. This possibility directly contradicts one of the AGM axioms of contraction.

More fundamentally, the AGM theory recognizes that there might be several different ways of revising beliefs, such as Rescher's preferred maximal consistent subsets, but nevertheless insists on a deterministic rule for producing a single resultant set of beliefs, either by selecting a single one of the possible revisions or by intersecting some or all of them to obtain the resultant beliefs. In the setting of practical systems for belief revision, however, it is natural to consider as well other forms for revision.

We broaden the study of entrenchment to consider the cases of indeterministic and probabilistic belief states as well as indeterministic selection of single revisions. In the enlarged conception, entrenchment becomes a modal concept that compares entrenchment in different extensions.

We say that $\Diamond[a \leq b]$ holds with respect to a set of base reasons m just in case there is some grounded configuration (m, s, x, y) such that the support of a depends on b. We obtain four possible states of affairs regarding each entrenchment in different extensions for each pair of beliefs a and b, as depicted in Figure 10: either (1) $\Diamond[a < b]$ and $\Diamond[b \not< a]$, (2) $\Diamond[a < b]$ and $\Diamond[b \not< a]$; (3) both $\Diamond[a < b]$ and $\Diamond[b < a]$ hold; or (4) neither $\Diamond[a < b]$ nor $\Diamond[b < a]$ hold.

The AGM postulates for belief revision concern only presence or absence of beliefs following a contraction or revision, with entrenchment constituting only a poor approximation of the notion of support. Are there corresponding postulates appropriate to characterizing revision in terms of grounded configurations? One might consider the intersection of some or all of the grounded configurations in an indefinite belief state as the result of rational revision, but this leaves beliefs without common support incomparable with respect to entrenchment.

What then are principles for reducing modal entrenchment relations to simple entrenchment? If the reduction is to be rational, then the preferential nature of reasons and criteria for identifying possible revisions suggests the Arrow criteria for group decisions, and as pointed out by Doyle [17] a version of the Arrow impossibility theorem for social choice then applies, showing that there is no general rational reduction method. This means that one can collapse modal entrenchment relations to simple ones only if the decision is made dictatorially, by a global comparison lacking in practice

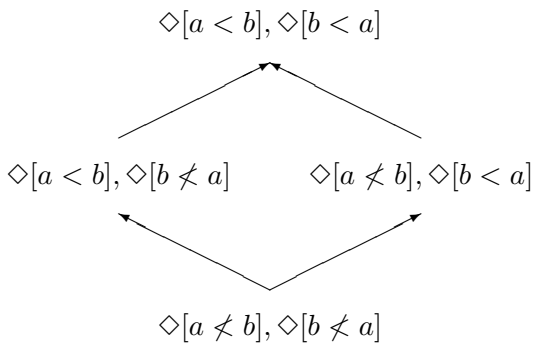

Figure 10: Fourfold states of possible entrenchment relations of two beliefs in indefinite belief states.

except on an ad hoc basis. In this case, one expects that different reductions will be characteristic of different kinds of agents.

6.2 Quantitative comparison via mechanical work

If one seeks a quantitative measure of difficulty of change, it does not help to simply choose a quantitative representation ρ of the entrenchment order such that $\rho(a) \leq \rho(b)$ iff $a \leq b$, for this would be a purely ordinal measure indicating nothing about the magnitude of changes. If one has a reliable measure of cardinal mental utility, one can use that to grade changes according to gain or loss of utility. But both of these paths assume one has an order or measure in hand, rather than showing how to measure change difficulty from the start. We now turn to a quantitative measure of change that does not presume the possession of an equivalent measure, but instead derives a measure from the notion of mechanical work or power from the magnitudes of external and refractory forces, of inertia, and of constitutional rigidity or inelasticity.

Measures of the effort of belief change depend on the representation of mechanical states and motions by which one characterizes beliefs. In the present treatment, we examine only changes characterized in terms of grounded mechanical configurations as defined previously, and in fact simplify the treatment by considering mechanical configurations to be of the form $(m, x, \dot{x})$ rather than the full structure $(m, \chi, \dot{\chi})$, as the latter follows the same pattern. In this setting of reasoning mediated by reasons, we can calculate the work and effort expended in reasoning as follows.

The *power* $P_t = \dot{x}_{t+1} \cdot f_t$ exerted across interval $(t, t+1)$ is found in the inner product of the force acting across that interval with the velocity across that same interval. The differing temporal subscripts of velocity and force in this formula reflect the difference between leading forces and

trailing velocities. For finite-basis vectors we calculate the instantaneous power P_t to be

$$\begin{aligned} P_t &= \dot{x}_{t+1} \cdot f_t && (43)\\ &= \dot{x}_{t+1} \cdot (\dot{m}_t, \ddot{x}_t) \\ &= |(\dot{m}_t, \dot{x}_{t+1}\ddot{x}_t)| && (44)\\ &= |(\dot{m}_t, \dot{x}_{t+1} - \dot{x}_{t+1}\dot{x}_t)| \\ &= |(\dot{m}_t, \dot{x}_{t+1})| - |(0, \dot{x}_{t+1}\dot{x}_t)|, && (45) \end{aligned}$$

here using the norm that counts the number of 1's in a vector corresponding to the ordinary inner product of binary vectors. The work expended across some interval, therefore, is the integral (sum) of the power over that interval, and we may use this mechanical work as a measure of the mental effort of reasoning. One needs a different norm or inner product when considering infinite-basis vectors, for example, when considering an inherent logic that produces infinite sets of conclusions such as infinitely many logically equivalent restatements $\{\neg\neg\}^* b$ of a particular belief b.

We examine this formula in greater detail to understand how this measure of effort works in the context of reasoning. Different reasoners operate in different ways, and we find that the overall effort of reasoning varies accordingly.

One can divide reasoning models into two types that differ in the treatment of time. In what one might call the *internal time* model, one identifies instants with steps of reasoning, no matter how long separates these steps in the world at large. In what one might call an *external time* model, one regards steps of reasoning as separated by intervals during which the reasoner does nothing. Both of these models offer useful insights.

We consider the internal time model first. This model corresponds best to a notion of deliberate reasoning, in which every step of reasoning involves some change to memory or outlook. In (45) we see that the power expended across a basic unit of time is the change of mass and position minus a cross term $\dot{x}_{t+1}\dot{x}_t$ involving velocity at successive intervals. This cross-term vanishes in deliberate reasoning because normally one does not immediately retract a conclusion one has just drawn, or draw a conclusion one has just retracted; there would be no point to it. In this setting, therefore, we obtain the magnitude of the power expended by the step of reasoning by

$$\begin{aligned} P_t &= |(\dot{m}_t, \dot{x}_{t+1})| && (46)\\ &= |\dot{m}_t| + |\dot{x}_{t+1}|. && (47) \end{aligned}$$

In this case, the work of a step of reasoning just adds together the number of changes made in memory and attitudes, so the effort involved in a chain of reasoning steps consists of the cumulative number of changes made in memory and attitudes across the span of reasoning.

In the external time model, we regard steps of reasoning as exerting impulse forces on the reasoner, with the reasoner exhibiting inertial (force-free) motion between steps of reasoning. The "Simple Impulse" table in Figure 11 illustrates the application of a simple impulse spatial force akin to the internal-time model just discussed. This impulse expends an effort of $|z|$ in the time step in which it is applied, according to the preceding calculation. In the subsequent inertial motion, of course, the force f_t vanishes, and so by (43) the power vanishes as well, so the total effort expended in a chain of reasoning steps again equals the cumulative sum of the number of changes to memory and attitudes, with the inertial motions doing no work.

Inertial motion takes a cyclic form in the discrete space $\mathbb{D}$ due to the algebraic characteristic that $x + x = \mathbf{0}$. As Figure 11 indicates, inertial motion with velocity z starting from a position

Time step	Simple Impulse				Up and Down			
t	x	$\dot{x}$	$\ddot{x}$	P	x	$\dot{x}$	$\ddot{x}$	P
0	$\mathbf{0}$	$\mathbf{0}$	z	$\lvert z\rvert$	$\mathbf{0}$	$\mathbf{0}$	z	$\lvert z\rvert$
1	z	z	$\mathbf{0}$	0	z	z	z	0
2	$\mathbf{0}$	z	$\mathbf{0}$	0	z	$\mathbf{0}$	$\mathbf{0}$	0
3	z	z	$\mathbf{0}$	0	z	$\mathbf{0}$	$\mathbf{0}$	0
4	$\mathbf{0}$	z			z	$\mathbf{0}$		
Work		$\lvert z\rvert$				$\lvert z\rvert$		

Figure 11: Kinematic quantities ($x, \dot{x}, \ddot{x}$), power (p), and total effort in two forms of spatial impulse motion starting from rest at the origin location.

z thus traverses the trajectory $z, \mathbf{0}, z, \mathbf{0}, \ldots$. It is certainly not commonplace to think of reasoners as cycling the last set of conclusions in this way. In standard artificial intelligence mechanizations, one instead regards step of reasoning as changing the set of conclusions from one set to another and then leaving it there until the next step of reasoning, as in the internal-time picture of motion. Accommodating this expectation requires one to modify the simplistic picture of reasoning seen in the internal time model.

One easily obtains a more familiar picture of reasoning by regarding steps of reasoning as exerting two impulses, corresponding to the rising and falling edges of a pulse, as depicted in the "Up and Down" table of Figure 11. That is, the force of the first half of a step of reasoning changes the velocity so as to effect the desired change of position, and the force of the second half of the step of reasoning changes the velocity back to zero by simply reversing (repeating) the thrust of the first half. This produces a pattern of motion of start-stop steps separated by zero-velocity intervals. This start-stop pattern of forces is in fact the pattern of reason forces, in which the frictional force component $-\dot{x}$ provides the falling impulse. This does not involve twice the mechanical effort of the internal time and simple external time pictures, however, because the falling impulses, matched with zero velocities in (44), contribute nothing to the cumulative effort.

Note that mechanical work only measures the effort of making the change itself, and do not include any effort involved in evaluating the applicability of some reasoning habit, of searching for the appropriate inference to perform, or of deciding among alternative inferences, if these activities are not effected by means of reasoning steps themselves. If generating a force z requires effort $|z|$, for instance, then the Simple Impulse motion involves an additional effort of $|z|$, while the Up and Down motion involves an additional effort of $2|z|$. The effort associated with such activities, however, depends on the organization and realization of the mind. For example, there need be no effort involved in producing the falling impulse of Up and Down as this value is already available as the velocity. Or for another example, evaluating the applicability of a set of reasons by a serial scan of databases of reasons and conclusions likely involves more effort than by parallel evaluations conducted by reasons wired together like neurons or Boolean circuits.

7. The likelihood of belief change

We regard the belief change undertaken in response to some motivated change as probabilistically dependent on a decision about whether to accept the motivated change at all, in part, or in full; on a decision about how to accomodate the accepted portion of the motivated change; and on effort of effecting the change. We look for a measure of the likelihood of change in which partial changes are more likely than more complete ones, and easier changes more likely than more difficult ones.

In the acceptance decision, the individual decides whether beliefs should change given the motivations presented. We regard susceptibility to different motives to change as reflecting strength of preferences between content of and grounds for beliefs, as sketched in Section 4.3.2. In this view, the likelihood of choosing one belief over another depends not only on whether the first is preferred to the second, but how strongly it is preferred, with likelihood of change increasing with strength of preference. This view does not presume perfect rationality, in that it allows a nonzero chance that an individual will retain a less preferred belief to a more preferred belief, but it does not forbid perfect rationality either. In fact, one might simply regard these likelihoods as measures of preference strength. Acceptance decisions might take into account of how difficult a change will be to effect, but need not.

In the accomodation decision, the individual chooses changes to beliefs in light of the acceptance decision. Accomodation decisions are more clearly dependent on the difficulty of different accomodations, with accommodations that involve less effort to effect being more likely than ones that require more effort.

7.1 Probability of change

As noted earlier, one can derive complete graded ascriptions from probability distributions over complete binary ascriptions by defining

$$\beta^* grade(b) = \int Pr_\Phi(\{\phi \in \Phi \mid b \in bel(\phi)\}), \tag{48}$$

Some natural distributions are those employed by Carnap [6, 38] in his theory of probability as "degrees of entailment." The *counting* measure $m^\dagger$ corresponds to making a Laplacian assumption that every possible configuration is equally likely to occur. This measure thus gives every configuration equal weight, i.e., $m^\dagger(\Phi') = |\Phi'|$, so that the probability of holding a belief is just the percentage of configurations over the base beliefs in which it appears. The *specificity* measure m^* regards configurations as partial descriptions of all the sets of beliefs extending them, and weights states by how "specific" they are, expressed formally as the proportion $m^*(\{x\}) = 2^{-|x|}$ relative to $2^{|\mathcal{D}|}$ of the number $m^{*\prime}(\{x\}) = 2^{|\mathcal{D}|-|x|}$ of possible supersets.

These probabilistic relationships derive simply from the structure of the set of possible belief ascriptions to configurations, but one also can consider probabilities engendered by other factors, such as economic content and mechanical properties.

One form for probabilities dependent on economic content uses probabilities to characterize failures of preferential optimization. Rather than assuming that all changes and configurations are optimal with respect to preferences over beliefs, this path assigns the largest probabilities to the optimal changes but assigns nonzero probabilities to other changes that decrease in magnitude as the level of suboptimality increases. For realistic notions of probability of optimization, however, presumably the measure assigned would need to reflect the likelihood of actually arriving at config-

urations of different degrees of optimality. As a fundamental tenet of theories of limited rationality is that it is sometimes harder to see the true optima than nonoptima that are locally optimal, deriving such probablities from degree of preferability alone would seem to miss the mark.

A perhaps more fruitful approach would base probabilities of change on the work needed to effect them, with changes involving less work appearing as more likely than changes involving more work. For systems conceived of as reflecting notions of potential energy, such transition probabilities would be a function of difference of energy levels, as in methods of simulated annealing [37]. One can use a similar approach in the current development, in which one uses the amount of work needed to effect each change in place of potential energy differences in assigning probabilities.

We note that development of a reasonable measure of probability of different revisions can be turned around to define a probabilistic entrenchment relation, or probability of entrenchment.

7.2 Expected difficulty of change

With a measure of the probability of different possible belief changes, no matter how these probabilities arise, one can combine these probabilities with measures of the difficulty of changes to obtain a measure of the expected difficulty of different changes. We note that a probability distribution over possible changes induces a distributional belief state even in a definite state individual.

8. The logic of belief change

To accompany this plain mathematical framework of states and branching-time histories, we use a temporal, dynamic, and multimodal logic for expressing statements about how beliefs and other properties of mental states change over time. We begin with the modal logic called CTL*, or Computation Tree Logic* [27], which is a full branching time logic. In this section, we augment CTL* with predicates germane to describing different types of belief states of individuals. In later sections, we will extend the CTL* basis with additional modalities akin to the program modalities of dynamic logic for expressing properties of complicated belief states and of actions that change beliefs.

The simplest statements in CTL* involve Boolean combinations of the underlying predicates and functions of states, such as ones indicating that some belief is held or not held. These state-specific statements then may be used in two types of statements, *path formulae* which make statements of linear temporal logic about what happens in a particular path of events, and *state formulae* which make statements about the futures possible at a particular state.

Path formulae of CTL* involve the modalities $\mathsf{F}\,p$, meaning that p is true sometime in the future; $\mathsf{G}\,p$, meaning that p is always true in the future; $\mathsf{X}\,p$, meaning that p is true at the next instant of time; p until q, meaning that p is true until q is true; and $p\ \mathsf{B}\ q$, meaning that p is true before q is true. These linear modalities are related in familiar ways: $\mathsf{G}\,p$ is equivalent to $\neg\,\mathsf{F}\,\neg p$; $\mathsf{F}\,p$ is equivalent to true until p; and $p\ \mathsf{B}\ q$ is equivalent to $\neg[\neg p\ \mathsf{until}\ q]$. We can similarly define the derived modalities $\mathsf{F}^{\infty}\,p$, meaning that p is true infinitely often, and defined by $\mathsf{F}^{\infty}\,p \equiv \mathsf{G}\,\mathsf{F}\,p$, and $\mathsf{G}^{\infty}\,p$, meaning that p is true at almost all times, defined by $\mathsf{G}^{\infty}\,p \equiv \mathsf{F}\,\mathsf{G}\,p$. Nonmodal statements are path formulae, and Boolean combinations of path formulae are path formulae.

State formulae of CTL* involve the modalities $\mathsf{A}\,p$ means that the path formula p is true in all futures, and the statement $\mathsf{E}\,p$ means that the path formula p is true in some future. Nonmodal statements are state formulae, and Boolean combinations of state formulae are state formulae.

The semantics of CTL* are based on Kripke semantics in which the possible changes of belief state are captured in an accessibility relation. We write $\psi S_o \psi'$ to mean that ψ' can follow ψ when the action o is taken. For present purposes, natural case to consider is one in which the accessibility relation corresponds to the successor relation between mental configurations in a nondeterministic history, in which case the accessibility relation would relate configurations rather than states, with $\phi S_o \phi'$ meaning that ϕ' can follow ϕ when action o is taken. This is the appropriate interpretation when considering the modal entrenchment relations discussed earlier. Such accessibility relations would reflect both the notions of changes of grounded extensions and also conservatism or minimal-change principles.

With no knowledge of any special connection between beliefs and mental states, the only theorems of the logic of belief are theorems of plain CTL* itself.

To treat different individuals and groups, one would add to the language constants naming the individuals and groups, as well as predicates naming the different individual types, group types, and group membership relations.

The logic of definite configuration individuals thus augments the base logic CTL* with the new predicates *In* and *Out*.

The logic of indefinite configuration individuals augments the logic of definite belief with the additional modalities $Arguable$ and its dual $Doubtless$, and $Provable$ and its dual $Conceivable$.

The logic of distributional configuration individuals augments the logic of indefinite belief with arithmetic inequalities about probability values such as $Pr(b) > 0.5$, possibly augmented with probabilistic modalities corresponding to commonly used conditions, such as $Pr(b) = 1$ (certain), $Pr(b) = 0$ (certainly not), $Pr(b) > 0.5$ (probable), or $Pr(b) > Pr(\neg b)$ (more likely than not).

For belief revision that takes into account the types of sources, content, form, and governmental role of beliefs, individuals, and groups, one would add names for these predicates.

It might also be useful to extend the vocabulary with additional predicates to characterize inherent beliefs and logical relationships, preferences among beliefs and grounds, and properties of beliefs and arguments regarding their defeasibility or indefeasibility by known arguments or by arguments of specific types.

In the modal setting, one can augment the modal entrenchment statements considered earlier with other modal statements. For example, the notions of arguability and provability take this form. Following the pattern of dynamic logic, we might write $\langle m \rangle b$ to mean that b holds in some grounded extension of the base set m, that is, that b is arguable in m, and write $[R]b$ to mean that b holds in every grounded extension of m, that is, that b is provable in m.

We can apply the same form of expression to characterize revisions, writing $\langle +r \rangle \phi$ to indicate that ϕ holds in some configuration following the addition of r to m, $[+r]\phi$ to indicate that ϕ holds in all configurations following the addition of r to m, and interpreting $\langle -r \rangle \phi$ and $[-r]\phi$ similarly in terms of removal of r from m.

9. Conclusion

The preceding presents an initial survey of the structure of belief revision and some candidates for quantitative measures of effort. Some of these measures are based on the connective dependencies relating different beliefs, and some are based on notions of mechanical force and work.

There are many directions for future work to extend the structures identified here. One direction is to use these measures in comparing the difficulty of learning and unlearning. Humans exhibit ob-

vious asymmetries in finding some things (e.g., bad habits) easy to learn and much harder to unlearn. We belief that further analysis along the lines begun here might shed light on these asymmetries of difficulty. Another direction for investigation is the development of revision methods based directly on mechanical concepts rather than procedural or logical ones.

References

[1] C. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction functions and their associated revision functions. *Journal of Symbolic Logic*, 50:510–530, 1985.

[2] Kenneth J. Arrow. *Social Choice and Individual Values*. Yale University Press, second edition, 1963.

[3] Robert J. Aumann. Markets with a continuum of traders. *Econometrica*, 32:39–50, 1964.

[4] N. D. Belnap. How a computer should think. In G. Ryle, editor, *Contemporary Aspects of Philosophy*, pages 30–56. Oriel Press, Stocksfield, 1976.

[5] Andrei Bondarenko, Phan Minh Dung, Robert A. Kowalski, and Francesca Toni. An abstract, argumentation-theoretic approach to default reasoning. *Artif. Intell.*, 93:63–101, 1997.

[6] Rudolph Carnap. *Logical Foundations of Probability*. University of Chicago Press, Chicago, 1950.

[7] Johan de Kleer, Jon Doyle, Guy L. Steele Jr., and Gerald Jay Sussman. AMORD: Explicit control of reasoning. In *Proceedings of the ACM Symposium on Artificial Intelligence and Programming Languages*, pages 116–125, 1977.

[8] Alvaro del Val. Non-monotonic reasoning and belief revision: Syntactic, semantic, foundational, and coherence approaches. *Journal of Applied Non-Classical Logics*, 7(2):213–240, 1997. Special Issue on Inconsistency-Handling.

[9] Jon Doyle. A truth maintenance system. *Artificial Intelligence*, 12(2):231–272, 1979.

[10] Jon Doyle. A model for deliberation, action, and introspection. AI-TR 581, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, 1980.

[11] Jon Doyle. Some theories of reasoned assumptions: An essay in rational psychology. Technical Report 83-125, Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 1983.

[12] Jon Doyle. Reasoned assumptions and Pareto optimality. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 87–90, 1985.

[13] Jon Doyle. Artificial intelligence and rational self-government. Technical Report CS-88-124, Carnegie-Mellon University Computer Science Department, 1988.

[14] Jon Doyle. Constructive belief and rational representation. *Computational Intelligence*, 5(1):1–11, February 1989.

[15] Jon Doyle. Reasoning, representation, and rational self-government. In Zbigniew W. Ras, editor, *Methodologies for Intelligent Systems, 4*, pages 367–380, New York, 1989. North-Holland.

[16] Jon Doyle. Rational belief revision. Presented at the Third International Workshop on Nonmonotonic Reasoning, Stanford Sierra Camp, CA, June 1990.

[17] Jon Doyle. Rational belief revision (preliminary report). In Richard E. Fikes and Erik Sandewall, editors, *Proceedings of the Second Conference on Principles of Knowledge Representation and Reasoning*, pages 163–174, San Mateo, CA, 1991. Morgan Kaufmann.

[18] Jon Doyle. Reason maintenance and belief revision: Foundations vs. coherence theories. In Peter Gärdenfors, editor, *Belief Revision*, pages 29–51. Cambridge University Press, Cambridge, 1992.

[19] Jon Doyle. Reasoned assumptions and rational psychology. *Fundamenta Informaticae*, 20(1-3):35–73, 1994.

[20] Jon Doyle. Final report on rational distributed reason maintenance for planning and replanning of large-scale activities (1991-1994). Technical Report TR-97-40, ADA328535, Air Force Research Laboratory, July 1997.

[21] Jon Doyle. *Extending Mechanics to Minds: The Mechanical Foundations of Psychology and Economics*. Cambridge University Press, London, UK, 2006.

[22] Jon Doyle. Cognitive mechanics: Natural intelligence beyond biology and computation. In Jacob Beal, Paul Bello, Nick Cassimatis, Michael Coen, and Patrick Winston, editors, *Proceedings of the Symposium on Naturally Inspired AI*. Association for the Advancement of Artificial Intelligence, AAAI Press, November 2008.

[23] Jon Doyle, Yoav Shoham, and Michael P. Wellman. A logic of relative desire (preliminary report). In Z. W. Ras and M. Zemankova, editors, *Methodologies for Intelligent Systems, 6*, volume 542 of *Lecture Notes in Artificial Intelligence*, pages 16–31, Berlin, October 1991. Springer-Verlag.

[24] Jon Doyle and Michael P. Wellman. Impediments to universal preference-based default theories. *Artificial Intelligence*, 49(1-3):97–128, May 1991.

[25] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77(2):321–358, 1995.

[26] J. Michael Dunn. *The Algebra of Intensional Logics*. PhD thesis, University of Pittsburgh, 1966.

[27] E. Allen Emerson and Joseph Y. Halpern. "sometimes" and "not never" revisited: on branching versus linear time temporal logic. *J. ACM*, 33(1):151–178, 1986.

[28] Kenneth D. Forbus and Johan de Kleer. *Building Problem Solvers*. MIT Press, Cambridge, MA, 1993.

[29] Peter Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. MIT Press, Cambridge, MA, 1988.

[30] Peter Gärdenfors. The dynamics of belief systems: Foundations vs. coherence theories. *Revue Internationale de Philosophie*, 172:24–46, January 1990.

[31] Peter Gärdenfors and David Makinson. Revisions of knowledge systems using epistemic entrenchment. In Moshe Y. Vardi, editor, *Proceedings of the Second Conference on Theoretical Aspects of Reasoning About Knowledge*, pages 83–95, Los Altos, CA, March 1988. Morgan Kaufmann.

[32] Matthew L. Ginsberg. Multi-valued logics. In *Proceedings of the National Conference on Artificial Intelligence*, pages 243–247. American Association for Artificial Intelligence, 1986.

[33] Joseph Y. Halpern. Reasoning about knowledge: a survey. In D. Gabbay, C. J. Hogger, and J. A. Robinson, editors, *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 4, pages 1–34. Oxford University Press, 1995. http://www.cs.cornell.edu/home/halpern/papers/knowledge_survey.pdf.

[34] Patrick J. Hayes. The frame problem and related problems in artificial intelligence. In A. Elithorn and D. Jones, editors, *Artificial and Human Thinking*, pages 45–59. Jossey-Bass, San Francisco, 1973. Reprinted in [58].

[35] R. C. Jeffrey. *The Logic of Decision*. University of Chicago Press, Chicago, second edition, 1983.

[36] Ralph L. Keeney and Howard Raiffa. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. John Wiley and Sons, New York, 1976.

[37] S. Kirkpatrick, Jr. C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, May 1983.

[38] H. E. Kyburg, Jr. *Probability and Inductive Logic*. Macmillan, New York, 1970.

[39] David A. McAllester. An outlook on truth maintenance. AIM 551, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, 545 Technology Square, Cambridge, MA, 02139, 1980.

[40] J. McCarthy and P. J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie, editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press, 1969.

[41] Drew McDermott and Jon Doyle. Non-monotonic logic—I. *Artificial Intelligence*, 13(1-2):41–72, April 1980.

[42] Robert C. Moore. Semantical considerations on nonmonotonic logic. *Artificial Intelligence*, 25(1):75–94, 1985.

[43] Walter Noll. Lectures on the foundations of continuum mechanics and thermodynamics. *Archive for Rational Mechanics and Analysis*, 52:62–92, 1973. Reprinted in [44].

[44] Walter Noll. *The Foundations of Mechanics and Thermodynamics: Selected Papers*. Springer-Verlag, Berlin, 1974.

[45] Walter Noll and Brian Seguin. Monoids, Boolean algebras, materially ordered sets. *Int. J. Pure Appl. Math.*, 37(2):187–202, 2007.

[46] W. V. Quine. *Philosophy of Logic*. Prentice-Hall, Englewood Cliffs, NJ, 1970.

[47] W. V. Quine and J. S. Ullian. *The Web of Belief*. Random House, New York, second edition, 1978.

[48] Willard Van Orman Quine. Two dogmas of empiricism. In *From a Logical Point of View: Logico-Philosophical Essays*, chapter II, pages 20–46. Harper and Row, New York, second edition, 1953.

[49] Raymond Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[50] Nicholas Rescher. *Hypothetical Reasoning*. North Holland, Amsterdam, 1964.

[51] Erik Sandewall. An approach to the frame problem, and its implementation. In *Machine Intelligence 7*, pages 195–204. University of Edinburgh Press, 1972.

[52] Leonard J. Savage. *The Foundations of Statistics*. Dover Publications, New York, second edition, 1972.

[53] Dana S. Scott. Domains for denotational semantics. In M. Nielsen and E. M. Schmidt, editors, *Automata, Languages, and Programming: Ninth Colloquium*, volume 140 of *Lecture Notes in Computer Science*, pages 577–613, Berlin, 1982. Springer-Verlag.

[54] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[55] Richard M. Stallman and Gerald J. Sussman. Forward reasoning and dependency-directed backtracking in a system for computer-aided circuit analysis. *Artificial Intelligence*, 9(2):135–196, 1977.

[56] Gerald Jay Sussman and Richard M. Stallman. Heuristic techniques in computer-aided circuit analysis. *IEEE Transactions on Circuits and Systems*, CAS-22(11), November 1975.

[57] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, third edition, 1953.

[58] Bonnie L. Webber and Nils J. Nilsson, editors. *Readings in Artificial Intelligence*. Morgan-Kaufmann, Los Altos, 1981.

[59] Michael P. Wellman and Jon Doyle. Preferential semantics for goals. In *Proceedings of the National Conference on Artificial Intelligence*, pages 698–703, 1991.

[60] Andrew W. Wicker. Influence models in social networks. Ph.D. proposal, North Carolina State University, defended April 8, 2010, February 2010.