UNIVERSITÉ PARIS CITÉ
École doctorale 474 - FIRE
Institut Jean-Nicod

Causal explanations and continuous computation

par

Can Konuk

Thèse de doctorat de Sciences Cognitives, Psychologie, Linguistique, Philosophie de la Pensée

Sous la direction du professeur Salvador Mascarenhas

Soutenue publiquement le 05/06/2025 devant un jury composé de:

1. Thomas ICARD, Professeur chercheur, Université de Stanford, Rapporteur
2. Jonathan KOMINSKY, Professeur assistant, Université d'Europe Centrale, Rapporteur
3. Alda MARI, Directrice de recherches, Institut Jean Nicod, CNRS, Présidente du Jury
4. Barbara HEMFORTH, Directrice de recherches, Université Paris-Diderot, Examinatrice
5. Ruth BYRNE, Professeur chercheur, Trinity College Dublin, Examinatrice
6. Tadeg QUILLIEN, Chargé de recherches, Université d'Édimbourg, Examinateur
7. Bob REHDER, Professeur chercheur, Université de New York, Examinateur
8. Salvador MASCARENHAS, Professeur chercheur, Ecole Normale Supérieure, Directeur de thèse.

Sous la direction du professeur Salvador Mascarenhas

0.1 Résumé court (français)

Titre: Formes logiques et explications causales

Résumé: Lorsqu'un événement a plusieurs causes, nous attribuons généralement différents degrés de responsabilité à chacune d'elles. Cette thèse étudie la manière dont ces jugements de responsabilité causale se rapportent à nos façons de représenter la causalité en général, d'une part, et à nos façons de représenter les probabilités, d'autre part. Certaines approches plus anciennes expliquent nos attributions causales en termes de forces d'association entre variables, généralement dans le cadre de modèles simples qui supposent peu de structure intermédiaire entre cause et effet. Plus récemment, les théories fondées sur les Modèles Causaux Structurels (MCS, ou Structural Causal Models, SCM en anglais) proposent des représentations plus riches, où des systèmes d'équations encodent des schémas de dépendance conditionnelle entre variables. Bien que les théories basées sur les SCM offrent de nombreux avantages, deux limites importantes peuvent être identifiées. Premièrement, les SCM mettent un accent trop exclusif sur les variables individuelles, considérées comme les constituants de base d'un modèle causal, ce qui les amène à négliger d'importants effets hyper-intensionnels dans la cognition causale: la forme logique dans laquelle une relation causale est représentée peut avoir un impact sur les jugements, un impact qui n'est pas saisi par la notion de dépendance contrefactuelle propre aux SCM. Deuxièmement, en privilégiant les probabilités plutôt que les forces d'association entre variables, ces modèles perdent leur lien naturel avec les théories de l'apprentissage, qui offraient une perspective sur la manière dont la connaissance causale et probabiliste peut émerger de l'observation répétée de co-occurrences entre variables. Dans cette thèse, je présente deux séries de travaux mettant en évidence ces limites. La première série d'études porte sur l'attribution de responsabilité causale pour des causes multivariées ("E est survenu à cause de A et B") dans le cadre des jugements dits de "sélection causale", où l'on identifie un sous-ensemble de causes d'un résultat observé comme étant "la" cause de l'événement, au sens commun. Ces études montrent de façon concluante que de telles conjonctions multivariées sont traitées par l'esprit comme de véritables entités causales et soulignent également combien la forme logique des règles causales peut influencer les jugements, une influence qui risquerait de passer inaperçue si l'on se concentre uniquement sur les causes singulières. Ensuite, j'examine comment les jugements de responsabilité causale influencent l'apprentissage causal des participants. Les résultats montrent, entre autres, qu'une explication en termes d'inférence bayésienne appliquée aux SCM ne parvient pas à rendre compte de certains schémas essentiels dans les données. Je soutiens que ces deux problèmes invitent à une solution commune. En m'appuyant sur la logique des programmes et sur des méthodes neuro-symboliques, je montre comment certaines relations causales traditionnellement décrites par les SCM peuvent être modélisées dans des architectures neuronales qui fournissent la structure supplémentaire nécessaire pour saisir les effets de la forme logique. Cette approche réintroduit la notion de force d'association, désormais distribuée au sein d'un réseau multicouche de relations, et permet de rendre compte de l'impact des explications sur l'apprentissage via des poids d'attention appliqués aux entrées.

Mots-clefs: Explications causales, Contrefactuels, SCM, Réseaux neuronaux, Modèles mentaux, Apprentissage, Théorie de la confirmation

0.2 Short abstract (English)

Title: Causal explanations and continuous computation

Abstract: When an outcome has multiple causes, people typically assign different degrees of responsibility to each cause involved. This dissertation investigates how such graded causal responsibility (or causal strength) judgments relate both to representations of causal structure and to representations of probabilities. Early approaches explained graded causal attributions in terms of association strengths between variables, typically within simple models. More recent theories, grounded in Structural Causal Models (SCMs), propose richer representations in which systems of equations encode patterns of conditional dependence among variables. Although SCM-based theories have successfully captured how people track counterfactual dependence, I identify two important limitations. First, SCMs encourage an overly narrow focus on individual variables as the basic moving parts of a causal structure by enforcing a one-equation-per-variable constraint and relying on a normative, possible-world semantics. This approach overlooks important hyperintensional distinctions: the logical form in which a causal relation is represented influences judgments in ways not accounted for by the SCM notion of counterfactual dependence. Second, by emphasizing probabilities rather than basic association strengths, these models lose the natural link to learning theories—an advantage held by earlier notions of causal power, which better explain how causal and probabilistic knowledge develops from repeated observation. I present experimental work that highlights these limitations. One series of studies examines causal responsibility attributions for multivariate causes (e.g., "E happened because of A and B") within causal selection judgments, providing conclusive evidence that such multivariate conjunctions are treated by participants as full-fledged causal entities by the mind and that the logical form of causal rules affects judgments in ways that a focus solely on singular causes would overlook. A second series of studies investigates how causal responsibility judgments impact causal learning, showing that accounts based on Bayesian inference over SCMs fail to capture key patterns observed in the data. Drawing on insights from program logic and neuro-symbolic methods, I propose that certain causal relations traditionally modeled by SCMs can instead be represented in neural models that incorporate an intermediate hidden layer between causes and outcomes. This approach reintroduces the notion of association strength—now distributed across a multilayer network—and elegantly captures the effect of explanation on learning via attention weights applied to inputs. For instance, an explanation such as "E happened because of A" directs learning by instructing the listener to focus on A when inducing a causal model for E.

Keywords: Causal explanations, Counterfactuals, SCMs, Neural networks, Mental models, Learning, Confirmation theory

Résumé

Cette thèse se propose d'étudier la manière dont nous appréhendons les relations de cause à effet en mettant en relation deux dimensions de notre connaissance causale: d'une part, la catégorie binaire de *cause*, par laquelle nous jugeons qu'un événement est la cause ou non d'un autre événement, et d'autre part, l'attribution graduée d'une *force causale* ou d'une *responsabilité* qui exprime l'importance relative d'une cause dans la survenue d'un effet. L'objectif est de montrer que nos représentations internes des relations causales possèdent une structure intrinsèquement graduée, laquelle se reflète aussi bien dans les explications que nous produisons au sujet des événements que dans les inférences que nous tirons des explications que nous recevons. Ce fait éclaire d'une part comment nous choisissons les explications à apporter à un phénomène lorsque celui-ci a plusieurs causes et d'autre part, comment les explications que nous recevons peuvent participer à l'acquisition de nouvelles connaissances causales. Je m'intéresse en particulier aux explications de *sélection causale* (causal selection) par lesquelles nous désignons parmi toutes les causes d'un phénomène un certain facteur particulièrement important comme étant *la cause* de celui-ci.

Dans la partie I, on analyse la manière dont les individus sélectionnent une cause ou un groupe de causes pour expliquer un événement donné (*causal selection*). Par exemple, lorsque la foudre tombe sur une forêt et y déclenche un incendie, pourquoi souligne-t-on la foudre plutôt que la présence d'oxygène, alors que ces deux conditions sont objectivement nécessaires à l'incendie? La littérature existante met en relief deux facteurs déterminants dans la sélection causale: la *normalité* (statistique ou prescriptive) d'un facteur d'une part et la règle causale qui lie l'ensemble des facteurs au résultat qu'il s'agit d'expliquer. Elle révèle par exemple que lorsqu'un événement a plusieurs conditions nécessaires, c'est souvent la plus *anormale* qui est jugée déterminante (*abnormal inflation*), alors que dans des contextes où plusieurs causes sont chacune suffisantes, on valorise au contraire la cause la plus normale (*abnormal deflation*). Ces effets ont été répliqués dans divers domaines, du jugement moral (Knobe and Fraser 2008) à l'explication scientifique (Morris et al. 2019), et sont robustes face à la manipulation de la consigne et du contexte étudié.

L'étude de causes "plurielles". Un apport nouveau présenté dans cette partie est l'exploration de *causes plurielles*, c'est-à-dire le fait que les gens peuvent aussi expliquer un événement par l'occurrence conjointe de plusieurs facteurs. Alors qu'une abondante littérature s'est focalisée sur l'attribution à un unique facteur, la thèse montre que nous sommes aussi capables de considérer des conjonctions de causes (comme " A *et* B ") et de les traiter comme candidats unifiés pour l'explication. Nous étudions les jugements de responsabilité causale produits par les participants à travers une séries d'expériences. Dans celles-ci, nous contrôlons à la fois la normalité des événements individuels — qui consistent en un tirage au hasard de boules de couleurs issues d'urnes qui contiennent chacune différentes proportions de boules de couleurs — et la règle qui relie

lesdits événements à un résultat d'intérêt (gagner ou perdre un certain round d'un jeu de hasard. Nous étudions ensuite comment les participants jugent de la *responsabilité* de telles conjonctions dans différents scénarios. Les résultats indiquent que la contribution d'un conjonction comme " A *et* B " est estimée par les sujets d'un façon holistique, qui ne se laisse pas réduire à une somme ou moyenne des contributions individuelles de A et de B.

Pour rendre compte de ces effets, je présente des théories contrefactuelles (Icard, Kominsky, and Knobe 2017; Quillien and Lucas 2023) qui reposent sur l'idée que, pour évaluer l'importance d'un facteur, les individus explorent mentalement des scénarios où ce facteur varie, tout en maintenant les autres à leur état réel ou à des valeurs jugées similaires, puis attribuent à chaque facteur une importance d'autant plus grande que faire varier ce facteur fait également varier le résultat considéré. La thèse présente comment ces théories, qui rendent déjà compte d'un grand nombre de motifs de jugements pour des variables individuelles, peuvent être étendues à des groupements composés de plusieurs variables pour rendre compte des résultats observés dans nos expériences.

Vers des modèles neuronaux avec structure interne. Dans ce chapitre, je défends l'idée qu'une explication fine des jugements de sélection causale exige de dépasser les modèles causaux structurels (SCM) traditionnels et de faire appel à des représentations neuronales non minimales, c'est-à-dire comportant des couches cachées qui matérialisent explicitement les conditions suffisantes minimales d'un résultat.

Je montre d'abord, à l'aide d'exemples simples, qu'une même fonction booléenne — par exemple

$$E = (A \wedge B) \vee (C \wedge D)$$

— peut être réalisée par plusieurs réseaux. L'architecture " disjonctive " que je propose réserve un neurone caché à chaque conjonction minimale ($A \wedge B$ et $C \wedge D$), ce qui conserve l'information sur la manière dont les facteurs se combinent et explique la préférence empirique des sujets pour les explications qui "restent du même côté" de la disjonction. Afin de donner un contenu computationnel précis à cette hypothèse, j'ancre les représentations dans le cadre des programmes logiques déclaratifs: chaque clause

$$E \leftarrow A, B$$

ou

$$E \leftarrow C, D$$

décrit une voie de preuve alternative. Ces programmes privilégient les conditions suffisantes positives et font apparaître la négation comme échec de la preuve ($\sim E$). J'exploite ensuite le résultat de neuro-symbolic computing connu sous le nom de CILP: tout programme logique général peut être traduit en un réseau de neurones à trois couches dont les poids et biais sont entièrement déterminés par la structure du programme. Cette traduction assure une correspondance formelle entre preuve logique et propagation d'activation continue. Sur ce substrat, j'introduis un processus de sampling contrefactuel de type MCMC: partant de l'état réel, l'agent propose de petites mutations (changement d'un seul facteur); leur acceptation dépend (i) de la normalité de l'événement, encodée dans un biais sur chaque neurone d'entrée, et (ii) de la cohérence structurelle: on décourage les transitions qui inversent l'activation d'un neurone caché. Cette dynamique suffit à reproduire l'ancrage empirique des contrefactuels "près du monde réel" et le regroupement spontané des variables A, B d'un côté, C, D de l'autre.

À chaque état ainsi visité, les poids sont ajustés par une procédure de Layer-Wise Feedback Propagation (Weber et al. 2025): le neurone-résultat reçoit une récompense, redistribuée couche par couche proportion-

nellement aux contributions effectives. Il en découle deux phénomènes classiques: (1) deflation des facteurs normaux aux poids `hidden`→`output` (ils se partagent l'explication), (2) inflation des facteurs anormaux aux poids `input`→`hidden` (ils deviennent des "leviers naturels" de l'issue). Enfin, j'utilise une mesure de Layer-Wise Relevance Propagation (Bach et al. 2015) pour "lire" l'importance causale directement dans les poids mis à jour; j'y ajoute un coefficient de complexité de chemin qui pénalise les explications mobilisant plusieurs voies disjointes, ce qui rend compte aussi bien de la parcimonie attendue pour les issues positives que de la prolixité observée lorsqu'il s'agit d'expliquer un échec ($\neg E$), lequel, dans le modèle, réclame une clause supplémentaire

$$\neg E \leftarrow\sim E.$$

Ainsi, le chapitre établit un pont formel entre programmes logiques, réseaux neuronaux continus et jugements causaux gradués, tout en unifiant sélection causale et explication des biais contrefactuels au sein d'un même cadre computationnel.

La partie II s'intéresse à la *réception* des explications, c'est-à-dire à la manière dont une simple explication "X est la cause de Y" qui se contente de désigner un ensemble de facteurs comme particulièrement important, peut guider l'apprentissage de relations causales plus complexes.

Tâche d'abduction causale. Pour étudier cette question, un paradigme expérimental innovant est présenté, dans lequel les participants doivent deviner la règle causale qui régit les gains et pertes d'un jeu de hasard à partir d'observations limitées: le joueur tire au hasard des boules provenant de quatre urnes contenant chacune différentes proportions de boules blanches et colorées, et observe s'il gagne ou perd à chaque tirage. Le résultat est déterminé par le tirage en fonction d'un règle que le participant doit deviner. Dans cette configuration, les participants ne peuvent pas observer toutes les combinaisons possibles de boules ni multiplier les essais; par ailleurs, le nombre de règles possible mettant en jeu le tirage des quatre urnes différentes est très vaste (de l'ordre de plusieurs dizaines de milliers de règles possibles a priori), ce qui rend l'inférence causale particulièrement complexe. Dans certaines conditions, des d'explications *sélectives* fournies sur ces mêmes observations, qui désignent un sous-ensemble de boules comme étant la cause du résultat observé. Les résultats montrent que recevoir une explication sélective (" vous gagnez *à cause de* A ") améliore notablement les performances d'inférence, par rapport à des conditions de contrôle dépourvues d'explications ou munies d'explications moins sélectives, qui mentionnent des variables pertinentes mais pas les plus cruciales . Nous avons également constaté que les sujets qui bénéficient de telles explications se forment des hypothèses plus focalisées sur les variables mentionnées, renforçant l'idée d'une fonction d'"orientation" de ces énoncés. Cela suggère que, dans des tâches où les variables observables sont nombreuses, le fait de pointer un facteur *en particulier* permet d'orienter l'attention des apprenants vers ce qui est véritablement pertinent.

Modélisation et portée théorique. Afin de clarifier les mécanismes cognitifs mis en jeu, nous avons formalisé la tâche au moyen d'un modèle bayésien qui repose sur l'idée que l'explication permet au récepteur de *reconstituer*, par inférence pragmatique, la théorie causale que le locuteur a en tête. Le modèle calcule pour chaque règle compatible avec les observations la probabilité qu'un locuteur aurait produit l'explication entendue dans l'hypothèse où il tient ladite règle pour être la règle qui régit le jeu. Il permet d'expliquer un certain nombre de motifs clés dans notre jeu de données.

Interprétation théorique des résultats. Je critique l'approche classique des ces résultats, qui consiste à dire que l'explication permet au récepteur de *reconstituer*, par inférence pragmatique, la théorie causale que

le locuteur a en tête. Toutefois, la thèse met en évidence les limites de cette approche: dès que le nombre de causes potentielles et la complexité de la règle augmentent, ce *reverse-engineering* inférentiel devient vite intractable pour l'esprit humain et ne rend pas compte du fait que nous apprenons *plus facilement* avec une explication, et non pas plus difficilement. L'alternative défendue est que les explications facilitent l'acquisition de connaissance causale en attirant l'attention de l'apprenant vers les facteurs les plus importants. Cette notion d'attention peut être articulée clairement dans le cadre d'un modèle connexioniste. Mentionner " X " revient dans ce cadre à amplifier l'activation du neurone X correspondant dans le réseau, ce qui biaise le processus d'ajustement des poids lors de l'apprentissage supervisé (ou par renforcement) et oriente plus rapidement vers un modèle dans lequel X est crucial. Il ne s'agit plus de calculer des inférences logiques complexes, mais de *restreindre* la recherche dans l'espace des solutions plausibles, en tirant parti de l'allocation d'attention. Je montre comment un modèle qui opérationalise cette notion d'attention permet de rendre compte des inférences faites par les sujets dans notre expérience mieux qu'un modèle bayésien basé sur l'idée de reconstruction par inférence pragmatique.

Conclusions générales et perspectives. Les travaux réunis dans cette thèse montrent:

1. Que les jugements de sélection causale reposent sur une représentation *graduée* de la responsabilité, sensible à la normalité et à la structure logique de la règle; les expériences menées soulignent l'influence de la normalité et de la structure causale, ainsi que l'existence de regroupements (causes plurielles) qui modulent fortement nos explications.

2. Que ces jugements, lorsqu'ils sont communiqués sous forme d'explications sélectives, guident efficacement l'apprentissage en restreignant l'espace de recherche plutôt qu'en imposant des inférences supplémentaires; des explications sélectives peuvent orienter l'attention d'un apprenant et faciliter l'acquisition de règles plus complexes, sans qu'il soit nécessaire de recourir à des inférences trop élaborées.

3. Qu'un formalisme unifié – où les poids d'un réseau portent à la fois l'information causale et probabiliste – offre un pont prometteur entre modèles symboliques et neuraux. Ces travaux ouvrent la voie à une articulation plus fine entre modèles symboliques et architectures neurales pour mieux comprendre et modéliser le raisonnement causal humain.

4. Une brève discussion est également esquissée sur la façon dont le même type d'approche pourrait servir à comprendre d'autres phénomènes, notamment dans le domaine de l'inférence probabiliste. Plus largement, ces résultats ouvrent des pistes pour l'élaboration d'outils pédagogiques ou décisionnels, capables de générer des explications adaptées à la fois à la structure causale du domaine et au profil attentionnel de l'utilisateur.

Acknowledgments

First and above all, I would like to address the first of these thanks to **Salvador Mascarenhas**, who has been my advisor during these three years of PhD, and before that as a master's student. One aspect of my debt to Salvador is obvious, and owes to his advice throughout these years as well as his direct contribution to some of the works presented in this dissertation. Another, less visible but for that reason only more worth mentioning, owes to Salvador's distinctive intellectual and conversational style – at once polemic and playful, and constantly enthusiastic for new ideas. I learned at his contact how not to be overly impressed by established theories, and how often their apparent simplicity rests on ignoring rather than solving puzzles. It also served as a constant reminder of the excitement and curiosity that make research worthwhile, and makes even risky ideas worth pursuing.

Another series of thanks should go to every member of my **dissertation committee**, who not only took the time to read it but also presented me with incisive, far-reaching questions to broaden my thinking on many of the questions raised in this work. This was so not only during the thesis defense, but also throughout the many discussions I had the opportunity to have with several members on the day after.

Before that, my thinking was deeply shaped by the stimulating discussions — and, at times, direct collaborations — I was fortunate enough to have with other researchers. Three individuals deserve special recognition, as without their input this dissertation might have taken a distinctly different form. **Tadeg Quillien**, whose early influence cemented and deepened my interest in causality and causal judgment, and whose rigor and clear-headedness was instrumental in setting some of the projects presented here on a sound foundation. **Nicolas Navarre**, with whom I had the pleasure of collaborating on the experiments in the second part of this dissertation, as well as on other projects not presented here, and discussing many of the ideas in this thesis as well as their relations to other corners of human judgment. I also had the privilege of regularly discussing many of the ideas presented here with **Thomas Icard**, starting from a stage at which they were still very embryonic. Thomas' own work itself had a disproportionate impact on the trajectory of my research, influencing — in one way or the other — every single chapter in this dissertation.

This work also benefited from my interactions with a number of other researchers, including Emile Enguehard, Michael Goodale, Tom Wysocki, and many others that I had the opportunity to meet in the environment of Institut Jean Nicod or in the conferences abroad I had a chance to participate in.

Finally, the experience of these years would not have been the same had it not taken place in my city of Paris, surrounded by all of those I know and love, and supported in particular by my two parents, who have seen the premises of this work long before it began. My mind was only free to pursue these ideas because my heart was firmly grounded in in the company of those around me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this dissertation I look at the relation between our category of *cause*, on which rests our knowledge of what events count as causes for a certain outcome, and the graded notions of causal *strength* and *responsibility* that underlie our intuition that the contribution made by a certain event to an outcome is large or small. Because the events we see as most responsible for an outcome also tend to be those we deem most worth mentioning to explain why that outcome happened, I study this relation through the causal *explanations* that people give for events. In particular, I am interested in the kind of explanations whereby we single out certain causes as more important, even as our causal knowledge supports the idea that their relation to an outcome is in some sense symmetric to other causes that we see as less important — for example, when we see both causes as strictly necessary for the outcome to happen.

My central claim in this dissertation is very straightforward. I say that these graded patterns are observed because our internal representations of causal relations are themselves inherently graded. In this sense, I take a different focus than popular accounts that see as the most defining feature of causal relations what in them cannot be reduced to graded quantities. That is, discrete parenthood relations between variables irreducible to statistical associations, that can be captured in structural models and associated with directed graphs. I do not, however, say that humans' representation of causal relations has less structure than is commonly argued, quite on the contrary. My argument is in fact that the structural models by which humans' causal knowledge is usually captured typically underestimate the amount of structure in our representations, by encoding only as many variables and edges as are required to track all the causal relations that hold true of a domain. Instead I wish to argue that these representations are more complex because they are constrained, not just by the necessity to describe all that we can observe or intervene on, but also by the mental devices through which we concretely represent what we can observe and intervene on. This leads me to understand causal explanations in terms of *neural* models, in which the relation between observable causes and effects is mediated by an intermediate hidden layer of neurons, and where the relation between each layer and the next depends on continuous weights and activation functions. Explanations are generated by internally using such models to run simulations over a causal domain and assess causal responsibility in terms of the connection weights on the paths between a variable and an outcome.

My use of connectionist models is however different from the one that consists in training them on data to show how they can reproduce human performance on certain tasks. Though this approach does a great job of illustrating the learning capabilities of such models, it is too unconstrained to offer any real understanding of human cognitive competence. Instead I base my hypotheses about structures on existing, *symbolic* theories of mental representations that offer independent insights into how causal functions can be computed concretely.

This enables me to circumvent the indeterminacy problems that usually plague connectionist theories of cognition, while also arguing that the explanatory purchase of such theories can be greater than that provided by symbolic theories on their own. I illustrate this explanatory purchase by looking at the relationship between causal explanations and internal representations of causal knowledge from its two ends successively. I ask how the mental devices by which we mentally capture the cause-and-effect structure of a domain inform the judgments by which we credit different factors for an outcome. And I also ask how the causal explanations we are given for events informs our acquisition of causal knowledge about these events.

The first part of this dissertation looks at how our representations of causal relations shape our causal explanations. I present new experimental work that constitutes the first systematic study of token causal responsibility judgments involving *multivariate* causes, where more than one unique event at once is invoked as *the cause* of the outcome. This work highlights how such judgments are sensitive to details of the logical form of causal rules as they are presented to them, and also interact with *negative* outcomes in ways that are not easily accounted for by structural model descriptions of their causal knowledge. I then show how hypotheses assuming connectionist representations with richer structure provides a better account of these same facts.

The second part looks at how causal explanations shape our causal knowledge. Here also, I present new experimental work, that looks at how subjects fare in a task where they have to infer the causal rule underlying a limited set of observations, and how their performance is affected when they are given causal explanations on top of those observations. Based on the results of that experiment and additional theoretical considerations, I then highlight the limits of existing accounts of inference from explanations based on a notion of reverse-engineering. Those accounts propose that subjects learn from explanations by considering what causal theories a speaker must hold true for them to have produced a certain explanation, based on shared conventions as to what explanations are appropriate in the context of each theory. I propose an alternative view, which I argue can only be clearly articulated in relation to the gradient-based learning processes supported by connectionist models. It sees the role of explanations as that of giving instructions to a listener as to which variables they should focus on to infer causal theories from observations.

In the rest of this introduction, I present a certain number of central notions and theories that will be relevant throughout this dissertation, which allow me to restate the central questions in more precise ways.

Causal strength and responsibility. Causal strength refers to the degree of influence a cause is seen to exert on an effect. In some contexts it manifest as a difference in *effect size*. Two people are each adding fuel to a fire; one pours a small canister, the other a large barrel of gasoline. Here the second action has a measurably larger impact on the spread of the fire. In other contexts, causal strength aligns with the extent to which a cause is seen as contributing to the occurrence of a *fixed* effect. A patient's recovery might for example have been influenced both by the rest they've had and the drug they've taken. Yet we sometimes reckon that one of those made a stronger contribution than the other — where the intuition seems to be that it was more likely to bring recovery on its own. In these latter instances in particular, our notions of causal strength blend into intuitive judgments of *responsibility*. That is, our categorical notions of cause and responsibility (whereby we consider an event as a cause or not in an all-or-none way) is complemented by graded notions, whereby we assign different degrees of responsibility to each cause.

Type and token judgments. Just like categorical notions of causation, such graded judgments can be articulated at a type- or token- level. Type-level causation concerns general relationships between event kinds. For instance, one might judge that smoking holds a high degree of responsibility in causing lung cancer. This would be a population-level or generic statement about a causal tendency, which may perhaps be quantified

by the increase in cancer probability for smokers versus non-smokers. In contrast, token-level causation apportions causal responsibility to a particular case: given that a particular patient, John, developed lung cancer, to what extent are his smoking habits responsible for that outcome?

Type-level strength informs token-level attributions to an extent: if I take smoking to have a strong influence in general, I will be inclined to assign it a lot of responsibility in John's particular case as well. Specifics of the case may however shift responsibility assignments. Suppose for example John is a non-smoker but has been exposed to asbestos. Though I may consider that smoking is a more important factor at the population-level than asbestos, I might still think that in John's particular case, asbestos is the main culprit.

Structural vs. graded aspects. In each of these causal judgments, one can distinguish a structural component and a graded component. The **structural** component asks what is the pattern of relationship between two events or kinds. Are they causally related? And if so, which event is the cause, and which is the effect? The fact that causal judgments engage intuitions about structure is most evident in their *directionality*: however strongly I might associate smoking with cancer, that will not lead me to believe that cancer causes smoking. Similarly, I might notice a strong correlation between "having stained teeth" and "developing cancer," yet if I consider both to be effects of smoking — rather than one causing the other — then I will not assign stained teeth any responsibility in bringing about cancer.

On the other hand, the **graded aspect** asks: to what degree did each factor contribute to the effect? Even if smoking and asbestos share the same directionality (both point to cancer, not the other way around), I might assign them different strengths in their causal roles. In short, structure tells me *which* way causation flows, while gradation tells me *how much* each cause matters for producing the outcome.

The Structure/Statistics dilemma. This duality makes the notion of graded responsibility or causal strength a particularly acute instance of what Smolensky (1987) calls the "Paradox of Cognition" or "Structure/Statistics Dilemma". On one side, when considering causal structure, we are driven to "hard" notions, which are discrete and all-or-none in nature ("A is a cause of B" vs. it is not, or $A \to B$ vs. $B \to A$). On the other side, when attending to gradation we are driven to "soft" notions that appeal to statistics and intuitions of partial influence. To a large extent, we can understand different approaches to causal strength as different ways of grappling with this paradox. Below I briefly review some historical attempts through that lens.

On one end of a spectrum, *strength-first* approaches treat **association strength** as the primary notion, from which aspects of causal structure are derived. They have natural ties to theories of learning such as classical reinforcement learning (see, e.g. Rescorla and Wagner 1972), that try to infer causal structure from patterns of associations between variables.

On the other end, *structure-first* approaches, emphasize complex causal structures, typically captured by the framework of **Structural Causal Models (SCMs)** (Pearl 2000; Spirtes, Glymour, and Scheines 2000). These support complex patterns of counterfactual dependence between variables and express graded notions of responsibility in terms of the probabilities associated with such patterns.

1.1 Strength first: the ΔP Measure

One of the earliest and most influential ways to quantify causal strength uses the **Delta P (ΔP) measure**. Formally, for a binary candidate cause C and outcome E,

$$\Delta P \;=\; P(E \mid C) \;-\; P(E \mid \neg C).$$

This difference captures how much C changes the probability of E. If C's presence substantially raises E's probability, ΔP is positive and large; if C lowers E's probability, ΔP is negative. Empirical work in contingency learning showed that human judgments of causal strength often correlate with ΔP (see e.g., Perales and Shanks 2017, for a comprehensive review), suggesting that people track difference-making as a core cue to causation.

A key insight in psychology is that ΔP-style associations can be *learned* over time via simple error-correcting rules. The most famous such rule comes from the Rescorla–Wagner model (Rescorla and Wagner 1972) in classical reinforcement learning, which posits that each cue X possesses an associative strength V_X. On a given trial, if the outcome is λ (often $\lambda = 1$ if present, $\lambda = 0$ if absent) and the total predicted outcome is

$$V_{\text{tot}} \;=\; \sum_{X \in \text{cues present}} V_X,$$

then the change in strength for cue X is:

$$\Delta V_X \;=\; \alpha_X \, \beta \, \big(\lambda \;-\; V_{\text{tot}}\big),$$

where α_X and β are learning rates. The term $(\lambda - V_{\text{tot}})$ is the *prediction error*—how far the actual outcome is from the sum of predicted contributions of the present cues. A closely related analog in neural network theory is the **Widrow–Hoff delta rule** (Widrow and Hoff 1960), also known as the Least Mean Squares rule, which updates weights via

$$\Delta w_i \;=\; \eta \, (t - o) \, x_i,$$

where w_i is the weight from input i, x_i is the input activation, $(t - o)$ is the error between target t and current output o, and η is the learning rate. Both Rescorla–Wagner and Widrow–Hoff illustrate how *association strength* can be incrementally acquired from experience; in simple cases, the predicted association converges on ΔP. Such perspectives are typically frugal in structural assumptions. In the extreme case of *model-free* reinforcement learning, no structural constraints affect which cues are considered; any candidate predictor can be included, and it will develop a nonzero weight only if the data support it. Such models also do not bake in explicit independence assumptions. The notion "X is or is not a cause of Y" then emerges as a thresholded or derived statement: if V_X is sufficiently above zero, we might say X is a cause.

1.2 An intermediate perspective: Causal Power theory

Despite their advantages, paradigms based on ΔP show some limitations when it comes to accounting for human judgments' sensitivity to the surrounding causal context, beyond the mere pairwise covariation between variables (see e.g., Holyoak and Cheng 2011; Perales, Catena, and Maldonado 2007). These limitations have prompted approaches that supplement contingency measures with additional structure. Such approaches can be seen as a perspective intermediate between purely associationist paradigms and structural causal models. They allow for richer models while restricting the functional form to *parametric* functions that can support error-based learning of parameters.

A particularly illustrative example of such approaches is found in **noisy-logical models**, for example, the noisy-OR. Suppose we have multiple causes $C_1, \ldots, C_k$ that can each independently bring about E with probabilities $w_1, \ldots, w_k$. The noisy-OR specifies:

$$P(E \mid C_1, \ldots, C_k) \;=\; 1 - \prod_{i : C_i \text{ present}} \big(1 - w_i\big).$$

This imposes a clear causal structure. Each C_i is a direct parent of E, whose generative contribution is assumed to be independent from others, ruling out interactions with the other causes. Such models go hand in hand with a precise notion of causal power. Cheng (1997)'s **Power PC theory** leverages this form: it argues that under assumptions of independence, the *causal power* of C_i can be expressed in terms of w_i, typically derived as

$$\text{Power}(C_i) = \frac{P(E \mid C_i) - P(E \mid \neg C_i)}{1 - P(E \mid \neg C_i)}.$$

This formula generalizes ΔP by normalizing out the baseline probability of E in the absence of C_i, thus reflecting how much C_i can produce E when other factors have *not* already produced it. The parametric constraints on noisy-logical models make it possible to learn each weight w_i from data via associative methods (Danks, Griffiths, and Tenenbaum 2003; Yuille and Kersten 2005).

1.3 Structure-first: graded strength through counterfactuals.

At the far end of the continuum, we find accounts that define graded strength in terms of patterns of counterfactual dependence between variables. These patterns are themselves taken to depend on rich models of causal relations, typically represented in the framework of **Structural Causal Models (SCMs)** (Pearl 2000; Spirtes, Glymour, and Scheines 1993).

An SCM is defined as a tuple $\mathscr{M} = (U, V, F)$, where:

- U is a set of **exogenous variables**, not caused by any other variable in the model.
- V is a set of **endogenous variables**, each of which is determined by other variables within the model (and possibly U).
- F is a set of **structural equations**. For each endogenous variable $V_i \in V$, there is a single equation:

  $$V_i := f_i(\text{Pa}(V_i), U_i),$$

  where $\text{Pa}(V_i)$ are the parent variables of V_i and $U_i \subseteq U$ represents the relevant exogenous influences.
- Exogenous variables are not determined by structural equations but are treated as random variables or error terms. In models dealing with **binary** variables, each exogenous variable $U_j \in U$ is seen as sampled from a Bernoulli distribution:

  $$U_j \sim \text{Bernoulli}(p_j),$$

  where p_j is the probability that U_j takes the value 1.

From the parenthood relations implicit in F, one can induce a directed graph $\mathscr{G}$, a bayesian network capturing the relationship between variables. SCMs typically come with the constraint that the graph $\mathscr{G}$ they induce should not contain any cycles, ensuring a well-defined causal ordering. Critically, an SCM is *non-parametric* in that each f_i can be any function (linear, Boolean, or otherwise) as long as it respects the acyclicity constraint.

Interventions and counterfactuals. The combination of random variables in U and structural equations induces a joint probability distribution over all endogenous variables of the structural model. SCMs enable

explicit reasoning about three types of conditional queries (Bareinboim et al. 2022; Pearl 2000): observational queries (e.g., $P(Y \mid X = x)$), interventional queries ($P(Y \mid \text{do}(X = x))$), and counterfactual queries ($P(Y_x = y' \mid X = x', Y = y)$). To illustrate, suppose we are looking at the effect of a drug (X) on patient survival (Y):

An **observational query** $\big(P(Y \mid X = x)\big)$ would ask: "Among patients who *chose* to take the drug, what fraction survive?". It would track *associations* in the observed data, without controlling for confounding factors that might affect both a patient's likelihood of taking the drug and their chance of survival (e.g. the severity of their symptoms). It arises from the SCM merely by observing the probability over states as naturally induced by the exogenous variables and endogenous functions.

An **interventional query** $\big(P(Y \mid \text{do}(X = x))\big)$ asks: "If we *assign* the drug at random to patients (i.e., intervene to set $X = x$), what fraction survive?". It aims to measure the direct causal effect of the drug on survival, unconfounded by factors that influence the choice to take the drug. This is what we do in the context of a randomized trial in which X is set in a way that is uncorrelated to its natural causes of occurrence. But we can also think of it as the result of a mental operation of fixing X to a certain value by *fiat*.

In an SCM, an intervention $\text{do}(X = x)$ is represented by *removing* the structural equation for X (i.e., we no longer let X be determined by its parents) and *replacing* it with a constant assignment $X = x$. This is equivalent to removing incoming edges in the corresponding graph $\mathcal{G}$. All other structural equations remain the same, and we then compute the new distribution of Y in this altered model.

A **counterfactual query** $\big(P(Y_x = y' \mid X = x', Y = y)\big)$ asks: "Given that a particular patient *did not* take the drug ($X = x'$) and *did* survive ($Y = y$), would they have survived if they *had* taken the drug ($X = x$)?". It aims to assess the effect of an intervention given specific background conditions that come with already observed facts. It can be represented in a SCM by first using the observed data ($X = x', Y = y$) to *identify* which settings of the exogenous variables U are consistent with a certain factual outcome. Then we *modify* the structural equation for X to set $X = x$ by intervention, while keeping the exogenous variables fixed to the values that generated that outcome. Finally, we re-compute Y under this new hypothetical condition.

1.4 Graded Causation in SCMs

SCMs emphasize rich representational structure, and in this sense they reverse the tendency exhibited by associationist models. In fact, part of the point of the framework is to highlight the sort of information that provably *cannot* be derived by contingency data alone (Bareinboim et al. 2022). This will be reflected in the sort of measures of causal responsibility that are defined over them. Graded causal responsibility in SCMs is often captured in terms of the probability associated with a certain counterfactual query. This is the logic behind measures like **Probability of Necessity** or **Probability of Sufficiency** (Pearl 1999, 2000). For example, Probability of Necessity (*PN*) roughly asks: "Given that Y occurred and X was present, how likely is it that Y would not have occurred had X been absent?" Formally, it can be written as

$$P\big(Y_{x'} = 0 \;\big|\; Y_x = 1\big),$$

where x' is a different value from x. Unlike parametric measures like (like ΔP or Power-PC) for which causal strength maps onto a local parameter of the causal model (like a weight on $X \to Y$), PN and PS quantify *emergent probabilities* that depend on the entire causal structure, including confounders, mediators, and exogenous variables U. They flip the associationist paradigm by making a continuous notion of strength emerge from a structural notion of Necessity. Whether X is necessary to Y can be reduced to a yes/no question at every given world, answerable based on structural equations alone. *How* necessary X is (or how sufficient, or how responsible) can then be cased out in terms of the respective probabilities of "yes" and "no" worlds,

also captured by the SCM.

1.5 Central claim and roadmap of the dissertation.

Having presented the main terms of the debate, I can lay out more clearly my primary position. I will introduce in Part I of this dissertation a specific class of token-level responsibility judgments, known as *causal selection* judgments, which will occupy center stage in this dissertation. These judgments powerfully support *counterfactual simulation* theories, in which the responsibility of A in causing B is viewed as a function of the covariation between A and B across counterfactual scenarios that subjects contemplate as they consider alternatives for what could have happened. This perspective is often closely aligned with the SCM framework, which elegantly encodes such scenario comparisons.

I will highlight the advantages of a such counterfactual simulation view, which I endorse. But I also argue that the choice SCMs as the machinery underlying counterfactual simulations is too limiting. Specifically, I'll argue that many of the patterns of responsibility judgments currently explained via structural models would be better understood by assuming different, *neural* representations of the same causal relations — representations that:

1. possess more structural details than SCMs,

2. break down that structure in terms of parametric functions with continuous-valued weights at the local level of pairwise connections between nodes.

In this sense, my approach borrows from both associationist (strength-first) and structure-first traditions. It sees the relational structure we see in a SCM as a high-level description of an underlying more granular network whose connections carry continuous weights. By presenting my attempt in these terms I am aware of the major difficulties associated with connectionist models of cognition more generally: such models can easily be too complex to be interpretable (see, e.g. Green 1998); their vast expressive power threatens to make connectionist accounts indeterminate with respect to any particular problem (see, e.g. Massaro 1988). There can, for example, be any number of networks representations compatible with the relation captured in the structural equation

$$E := (A \wedge B) \vee C. \tag{1.1}$$

This in turns puts our choice of model for how humans represent that relation at risk of being completely *ad hoc*. For these reasons many researchers contend that, although connectionist models may ultimately give more accurate pictures of what cognition looks like at some lower level, they do not by themselves contribute to the scientific understanding of higher cognitive functions. Instead, understanding can only come from the level of idealization embodied by structural models and often associated with Marr (1982a)'s "computational" level of analysis. Structural models provide a constrained symbolic framework, built on top of classical propositional logic that puts clear bounds on what sorts of models might capture the causal knowledge that we attribute to a given subject.

I acknowledge this concern but propose that it may be circumvented via a detour through an intermediate level of idealization, somewhere between fully unconstrained connectionist networks and SCMs. The boundaries of this intermediate level are defined by a family of theories that remain symbolic in nature, but that go beyond the *classical* representations found in most computational-level models. They draw on fields such as formal semantics of natural language (Groenendijk 2008; Mascarenhas 2009), mental model theory (Byrne 2005; Johnson-Laird 1983b; Koralus and Mascarenhas 2013) or declarative logic programming

(Lloyd 1984; Robinson 1965). Such algorithmic-level accounts have in common that while they traffic in representations with symbolic content, the sort of representations that they appeal to typically come with more constraints than the classical representations underlying accounts like SCMs.

For example, they provide grounds to argue that there is more to people's representation of the disjunctive rule in (1.1) than is captured by the semantics of classical disjunction. They highlight how humans would intuitively think of such rule, not as embodying a single causal mechanism (as encoding it in a single SCM equation would suggest) but in terms of two distinct routes for generating E, embodied in two alternative mental models $\{A \wedge B,\ C\}$ each tracking one minimal set of conditions to make E true, or in two separate clauses in a logic program:

$$\{E \longleftarrow A, B; \quad E \longleftarrow C\} \tag{1.2}$$

Translation algorithms inspired from the research on neuro-symbolic architectures (Garcez, Broda, and Gabbay 2002; Garcez, Lamb, and Gabbay 2009) will allow us to go from these intermediate representations to neural architectures which I assume implement them at a lower level. For this particular example —leaving detailed generalizations for later—, this will result in a three-layer network with continuous weights as in Figure 1.1 in which the hidden layer structure mirrors the extra structure posited by mental-model type accounts and which classical SCM accounts obscure. This synthesis avoids the underdetermination of purely connectionist accounts by tethering network architecture to constraints from pre-existing, symbolically articulated theories — which themselves reflect cognitive phenomena already captured at higher levels of analysis. Crucially, however, the neural framework does not merely "cash out" these symbolic insights at a lower level of implementation. Instead, it extends their explanatory reach, in ways that we'll explore in several stages through this dissertation.

In **Part 1**, I explain how by integrating these insights about human's representational arsenal with networks holding clear causal structure, we can account for patterns of human causal judgments that are hard to reconcile with SCM-based accounts. In particular, I present the first series of experiments looking at people's causal selection judgment for multivariate causes ("E happened because of A and B"). These will evidence patterns of judgments that defy traditional accounts, in particular illustrating the fact that people's judgments are sensitive to effects of causal structure that go beyond the individual variable, and do not seem by default to handle explanations for negative outcomes (explaining why "$\neg E$" happened) in classical ways.

After presenting these findings, I show how the same facts can be elegantly handled by assuming that people represented causal relations and generate counterfactual simulations over a neural model with the sort of structure outlined above. In a nutshell, the account will assume that people explore alternatives close to the present situation, and *increase* the weights associated with the neurons that most contribute to the outcome in these nearby states. The respective importance of each variable will then be a function of the magnitude of the weights on its path(s) to the outcome after such updates have been performed. This perspective effectively brings back the intuition found in associationist models that the causal strength underlying subjects judgments maps directly onto parameters of their internal models.

Part 2 then highlights the advantage of bringing back such associationist intuitions by looking at causal explanations from the angle of learning. To that effect I present a new experimental paradigm where subjects were tasked with guessing the causal rule underlying a set of data (observations of the shape: "A and B happened, C and D didn't, and in this configuration the outcome E happens."), which in some conditions were augmented with causal selection explanations (e.g., "E happened because of A"). I show that current account of how explanations fit into such learning tasks, which assume that explainees reverse-engineer the causal theory that explainers have in mind via Bayesian inference, fails to explain key patterns in the data. By contrast, I show how a connectionist model with very minimal assumptions can capture the role of explanations in causal learning very elegantly. Its core premise is that the role of an explanation like "E

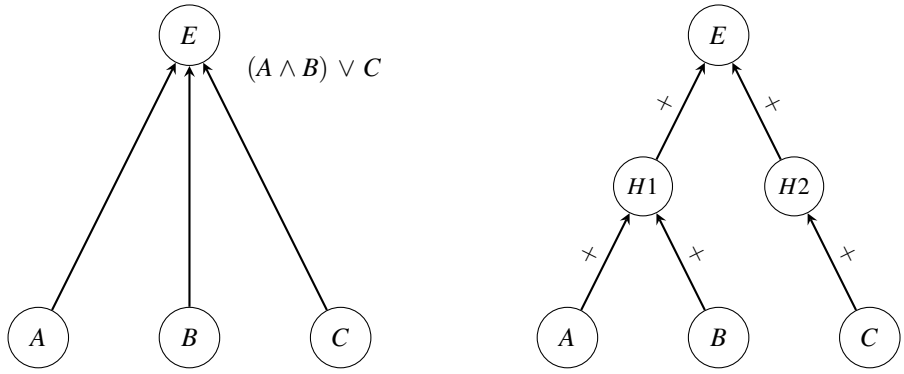


Figure 1.1: A schematic comparison of an SCM DAG (left) with an equivalent three-layer neural network (right). In the DAG, E is given by $(A \wedge B) \vee C$, meaning E depends on both A and B jointly, or on C. In the neural network, we introduce two hidden nodes: H1 handles the conjunctive route (*A and B*), while H2 handles C. With suitable biases and activation functions, H1 effectively "turns on" only if *both A* and B are active, whereas H2 is driven solely by C. The output layer E then combines (disjunction) these two internal signals. Each edge is marked with "+" to indicate a positive weight. This architecture mirrors the extra structure (two distinct routes) that a mental-model or clause-based perspective might attribute to $(A \wedge B) \vee C$, but cast in continuous, parametric form.

happened because of A" is to tell a learner what part of the input they should focus on as they engage in error-driven learning procedures. This can be modeled straightforwardly by complementing such procedures with an *attention* parameter that inflates the activation of neurons associated with the variables mentioned in the explanation (e.g. "A"), and deflates that of others. This in turns inflects the learning trends obtained as a result of backpropagating error signals through the model, in ways that favor theories giving more importance to the mentioned variables. Besides accounting for experimental data, this account also captures the strong pre-theoretical intuition that explanations do not make learning harder (by requesting additional reverse-engineering inferences from the learner) but easier, by alleviating the learner's *frame problem* of figuring out which variables are relevant to draw inferences from observations to begin with.

Part I

Causal selection judgments and causal representations

Chapter 2

Introduction to Part I

Causal selection is the process whereby, among the many causes of an outcome, we single out certain events as particularly prominent. In ordinary language, this manifests in the kinds of judgments whereby we say that "A is *the cause* of B" or that "B happened *because* of A".

A classic illustration (cf. Hart and Honoré 1985) is given by the following: suppose that a forest catches fire after having been struck by lightning. Intuitively, one wants to say that *"The lightning bolt caused the forest fire"* rather than assigning responsibility to (for example) the presence of oxygen in the air. This example clearly demarcates causal selection judgments from mere categorical judgments of actual causation (like those studied by, e.g. Halpern 2016a) in the sense that, from a categorical standpoint, the contribution of each factor is exactly symmetric, as both the lightning (L) and oxygen (O) are necessary conditions for the fire (F) to happen. One could capture their relationship with the conjunctive rule:

$$\mathtt{F} := \mathtt{L} \wedge \mathtt{O}. \tag{2.1}$$

Yet people see a clear difference in the degree of responsibility that each event possesses. Such graded intuitions are a stable feature of people's judgment in such cases, even when controlling for other possible confounds such as time-dependence or confidence in the existence of a causal link for each event (cf. O'Neill et al. 2021).

Abnormal inflation. Further studies (e.g. Icard, Kominsky, and Knobe 2017) reveal that a major driver of this asymmetry is the fact that the lightning bolt is a *rarer, more abnormal* event while on the contrary the presence of oxygen is extremely ordinary. Indeed, one of the most consistent patterns of causal selection judgment, known as *abnormal inflation* refers to the phenomenon by which, in a context where two or more variables are individually necessary for an outcome to happen, the most abnormal factors will be judged more causally responsible[1]. Note that the relevant notion of *normality* is not strictly statistical frequency. It can also include moral or conventional norms. Icard, Kominsky, and Knobe show that events that are abnormal from a moral or conventional point of view — even if not low-probability—will also receive greater causal credit.

In the rest of this dissertation – following a strategy adopted by much of the literature on causal selection– we will mostly focus on cases involving statistical normality, for the simple reason that it is easier to

[1] As Icard, Kominsky, and Knobe (2017) discuss it, "abnormal inflation" is strictly about inflating the credit of a variable as it gets less normal. Throughout this dissertation I will use the notion in a looser sense, that lumps together this effect and the related phenomenon of "supersession", which is about inflating the credit of a variable as the alternate candidate cause becomes more normal.

manipulate experimentally than prescriptive norms. But we keep in mind that the relevant notion of normality behind causal selection judgments ultimately involves an amalgam of statistical and prescriptive intuitions.

Abnormal deflation in disjunctive cases. Importantly, the preference for *abnormal* variables is not systematic but depends on the kind of causal relation that variables entertain with the outcome. In a situation where two or more variables are each individually *sufficient* for an outcome to happen, people's intuitions tend to reverse the abnormal inflation pattern. Suppose for example that the plants in your garden are constantly and reliably watered by a system of sprinklers, in a region where rain is rare. On any given day, whether or not your plants will be watered (W) can be captured as a function of the activity of the sprinklers (S) and the occurrence of rain (R) via the disjunctive rule:

$$\mathtt{W} := \mathtt{S} \vee \mathtt{R}. \tag{2.2}$$

If now on a given day, the occasional rain fell while the sprinklers also did their usual job, the outcome W is *overdetermined* in the sense that either events alone would have ensured the plants are watered. In the presence of such disjunctive rules, people tend to exhibit the tendency *opposite* from abnormal inflation, in that they'll prefer to point to the normal, expected event (here the action of the sprinklers) as the cause of the watering of the plants. The phenomenon is termed **abnormal deflation** by Icard, Kominsky, and Knobe (2017).

The interplay of normality, structure, and facts. All in all, causal selection judgments arise from an entangled mixture of three factors: the normality of candidate causes, the causal structure linking events, and the actual events that transpired in a given situation. While abnormal inflation (preferring rare causes in conjunctive settings) and deflation (preferring normal causes in disjunctive overdetermination) provide paradigmatic examples of such effects because of their simplicity, the same entangled effects extend to more complex systems (see, e.g. Quillien and Lucas 2023, for a series of experiments dealing with more complex causal structures).

Contrastive effects in causal selection. Interestingly, such judgments seem inherently contrastive, in that our assessment that "E happened because of A" is sensitive to the salient alternative causes "B, C, ..." that we could have given as explanations in lieu of A. The following example from Dretske (1972) can serve to illustrate. Suppose that Clyde is the son of John, a rich magnate who stipulated in his will that Clyde can only inherit from him if he marries before 30. Clyde marries his girlfriend Bertha to get the inheritance. Now in that context, it seems that while sentence (1a) below would be an accurate statement, sentence (1b) would be false.

(1) a. Clyde got the inheritance because he MARRIED Bertha.
    b. Clyde got the inheritance because he married BERTHA.

The effect of such linguistic focus as is glossed by the caps in the examples in (1) is classically understood (Rooth 1992) in terms of the set of alternatives that it raises. While both sentences enunciate the same state of affairs, (1a) highlights the alternatives {marrying, not marrying}, whereas (1b) highlights {marrying Bertha, marrying someone else}. Suppose, in an alternative scenario, that Bertha was not Clyde's girlfriend, but instead Clyde is engaged to Sue. John, who disapproves of his son's choice of partner, conditioned his inheritance to the fact that Clyde does not marry Sue. In that new context, (1b) becomes the more attractive explanation while (1a) becomes plain false.

2.1 Counterfactual simulation models of causal selection

2.1.1 Causal selection and associationist theories.

Causal selection judgments are interesting (among other things) because they present a puzzle for associationist theories of causal strength. At first glance, a measure like ΔP (the difference in outcome probability with vs. without a cause) appears plausibly aligned with these judgments. They track the patterns of abnormal inflation and deflation qualitatively when normed over possible candidate causes (as shown in e.g. Quillien (2020)). And such norming is arguably warranted in view of the contrastive nature of causal selection judgments highlighted above. The underlying theory here would be that, in the forest fire scenario for example, people assign higher strength to lightning (L) than oxygen (O) because they observe across contexts that lightning strikes covary with fires a lot more so than with oxygen (which is ever-present, including in non-fire contexts), and thereby learn to associate the former with fires more strongly than the latter.

However, the facts highlighted in the previous section and elsewhere in the experimental literature reveal the limitations of such an account. First, causal selection often occurs in *single trial* contexts without repeated observations. Participants routinely make judgments (e.g., singling out A as "the cause" in $\mathtt{E} := \mathtt{A} \wedge \mathtt{B}$.) after learning a causal rule *once*– no repeated observations are given. Often also vignettes involve events sufficiently abstract that people can't be suspected to leverage previously acquired world knowledge (see e.g. Morris et al. 2019). Second, we've seen that causal selection judgments are sensitive to *prescriptive* norms, which by definition cannot be directly observed. Third, responsibility attributions pivot on context-specific occurrences, in ways that are hard to explain in terms of association strengths that track the dependence between variables *across contexts*.

For all these reasons, it would seem that the sort of covariation that is relevant for explaining patterns of causal selection is not so much the covariation between events as they occur in the real-world, but instead their covariation across scenarios as those *occur to us* in mental simulations. This is the central idea behind counterfactual simulation models of causal selection judgments, which I to develop in more details in the next section.

2.1.2 Counterfactual Models: An Analogy and Key Ideas

Perhaps the easiest way to grasp the idea behind counterfactual simulation models is by analogy to the way in which we track down the causes of an event as we interact with a real-world causal system. Suppose your computer is lagging. You want to understand why. Perhaps you have too many tabs or applications open? You close some to see if that changes anything. Perhaps the machine is overheating? You lift it off your desk to let fresh air circulate behind it. Perhaps the files you are downloading are too large, so you try pausing the download, and so on. After each manipulation, you can assess what effect it has on the lag of your computer, which will give you a notion of how much each factor was responsible for your computer's slow performance.

Note that in all of these manipulations you are not so much trying to understand how your computer works in general – in fact, your choice of interventions is driven in large part precisely by the fact that you already have some pre-existing knowledge on its workings. Rather, you are trying to pin down which particular factor is responsible for the lag in *this* particular case. Throughout the procedure, you engage in these successive steps in an order that is *non-random*, but clearly determined by at least two factors.

First, you are guided by the readiness with which each of these possible manipulations *occur to you* to begin with, owing to some context-dependent saliency effects. Closing tabs might occur to you first, for example, because you are currently working with them, while the app you opened yesterday, while a plausible culprit in principle, might come to mind only later.

Second, the manipulations you can engage in, as well as the consequences that they have, are also determined by some "hard" structural features of the system you are interacting with. Some variables can't be manipulated without affecting others. You may think for example that you have signal issues, and consider rebooting your network connection. But this cannot be done without *also* interrupting your current download, which would confound the diagnosis. Structural features also impose constraints on the granularity of your manipulations. If you are currently downloading a folder of ten files, you may not (under some conditions) be able to interrupt the download *only* for files 3 and 4. The action has to scope over the download as a whole.

In this example, the causal system is an *external* one (the computer). Now counterfactual simulation theories essentially submit that we do something analog *internally* as we use our causal knowledge to generate an explanation for some event occurrence. Essentially, we engage in a two-step process.

1. We run *counterfactual simulations*, whereby we vary some elements of the situation that we've observed (and seek to explain) to explore certain alternative ways things could have been. Just like our decisions as to which manipulations to engage in are determined by some context-dependent saliency as we interact with our computer, here our decisions as to which alternatives to contemplate will be determined by factors such as the *normality* associated with these scenarios (normal events more readily come to mind than abnormal, a fact independently evidenced in e.g. Byrne (2016) and Kahneman and Miller (1986)). It will also be anchored to *real-world* occurrences of events, in that we are more prone to contemplate alternatives that are similar to the real-world (Lewis 1973a; Lucas and Kemp 2015).

2. In each such simulation, we use our internalized causal knowledge to generate the outcome that we expected to follow in this new, imagined context. The degree of responsibility that each cause possess can is then seen roughly as a function of how it covaries with the outcome across the set of all such simulations.

   Specific theories sometimes posit additional manipulations occurring in each simulation that refine the measure beyond mere covariation, that involve for example checking how the outcome as we manipulate the value of some focal cause in each counterfactual, to get a measure of counterfactual effect size (Quillien 2020), or assessing the correspondence between focal cause and outcome under different values of the surrounding causes, to track notions of Necessity and Sufficiency of that cause for the outcome (Icard, Kominsky, and Knobe 2017).

We will get into more concrete modeling details for two such counterfactual theories (Icard, Kominsky, and Knobe 2017; Quillien and Lucas 2023) in the work we present in Chapter 3 below. For the remarks I want to make presently, it will be enough that we grasp the general picture, by considering how such theories would apply to the forest fire example introduced previously. After observing that both as both the lightning (`L`) and oxygen (`O`) were present, and equipped with the knowledge captured by eq. (2.1), we proceed as follows:

1. Imagine a number of alternative scenarios to the present one; because normal scenarios come more readily to mind than abnormal ones, these will involve more scenarios where the lightning bolt didn't strike (`L`=0) but the oxygen was still present (`O`=1), than the converse.

2. Use our internalized knowledge to roll out the consequences expected in each of those scenarios. In the worlds generated in this way, it will be apparent that the presence of the fire correlates more closely with that of the lightning bolt, because every time the lightning bolt is present, the fire occurs (as the oxygen is almost always there to satisfy the other necessary condition), while the reverse isn't true.

Now, just like in the computer analogy above, it is apparent that all of these manipulations are sensitive to the **structure** of the program we are interacting with – in this case, the *internal* program whereby we track and simulate facts about forest fires. The manipulations whereby we construct alternative scenarios pivot around the event tokenings `L` and `O`, whose value we can change while keeping the rest constant. If we were to engage in more manipulations within each imagined scenario – as some theories assume we do –, such as bringing back the lightning bolt in a scenario from which it is absent (to see if this is sufficient to bring back the forest fire with it), those would also depend on a pre-existing structure to carve out the "moving parts" through which our operations can have a handle on the system.

From that point of view, a premise shared by most existing counterfactual simulation models of causal judgment is that Structural Causal Models – that is, in this case, the SCM built on top of eq. (2.1) – provide the appropriate structural scaffolding on which all of those manipulations can rest. While engaging with the overall spirit of simulation models, my argument throughout this chapter will consist in challenging this premise.

To make that argument concrete, one will have to look beyond the sort of simple examples provided here as illustration. This is what I'll do in the first part of this chapter, where I present a series of experimental studies, the result of joint work in collaboration with T.Quillien and S.Mascarenhas. That section reproduces verbatim a self-contained paper recently submitted to *Cognition*. These studies present the first systematic exploration of causal selection judgments involving multivariate causes like

(2) "E happened because of A and B".

We look at such judgments in contexts involving complex causal rules combining disjunctions and conjunctions, and also explore subjects' judgments in situations where they're tasked with explaining *negative* outcomes (i.e. explaining why one lost a game, after having been instructed in the winning conditions). These experiments provide convincing evidence that people engage with conjunctive causes (like $A \wedge B$ in (2)) as whole entities, assessing their causal responsibility profile holistically rather than reconstructing it from that of their component parts. They also show how subjects judgments depend on the logical form of causal rule in a way that isn't captured by classical SCM-type structures.

In that paper, we hypothesize that subjects' pattern of judgment can be understood by assuming that they operate with a non-classical set of representations as they engage in counterfactual simulations. We propose to understand these representations by analogy with natural language *plurals*, a concept we explained in more detail in the paper. This provides a symbolic-level description which, in the second, more theoretical Chapter 4 – written specifically for this dissertation – I propose to re-analyze these results at a lower, subsymbolic level by proposing that the internal manipulations subjects engage in as they generate explanations operate on a richer structure than that revealed by structural models.

Chapter 3

Plural Causes

Author note: This chapter is a verbatim reproduction of a manuscript currently under review, written in collaboration with Tadeg Quillien and Salvador Mascarenhas, with myself as first author (Konuk, Quillien, and Mascarenhas 2024).

3.1 Introduction

Causal selection is the process underlying our intuition that an outcome happened *because of* a given event, or that an event is *the cause* of an outcome. Causal selection judgments go further than judgments of *actual causation* (Halpern 2016b; Halpern and Pearl 2005; Hitchcock 2001b), whereby people merely identify which events can be counted as causes of an outcome. They induce a ranking over these events, singling out some as being more important than others in bringing about the outcome under consideration. When a forest catches fire after a lightning strike, for example, people tend to say that the lightning bolt was the cause of the fire, not mentioning the presence of oxygen in the air, although they are well aware that the latter was no less indispensable for the fire to occur. Causal selection in this sense is crucially distinct from *causal inference*, the problem of learning the relevant causal facts about the world. Causal selection concerns how we judge the relative importance of the many causes of an event, given that we already have a causal model of the situation.

A considerable literature has developed around what factors underly our preference for certain causal explanations of an outcome over others (Icard, Kominsky, and Knobe 2017; Knobe and Fraser 2008; Morris et al. 2019; Quillien and Lucas 2023). Although theories diverge as to what the drivers of causal-selection judgments are, they all agree that the outcome of causal-selection judgments depends crucially on the initial pool of candidates under consideration.

Before the lightning bolt can be viewed as *the cause* of the fire, the events *lightning*, *oxygen*, *dry season*, and others must first be flagged by the mind as relevant candidates for causal selection, whose relative importance in bringing about the outcome will be assessed. We argue here that the extant literature on causal selection has had a blind spot regarding that initial pool of candidates: it operates on the implicit premise that the only relevant variables for causal selection are *individual variables*, corresponding to distinct nodes in the relevant network of causes.

Instead, we argue that causal selection judgments can recognize *plural* causes, featuring more than one variable, as when we say that "the dryness of the season and the strength of the wind" caused the uncontrollable spread of the fire. We argue that such plural causes are treated by the mind as candidate

explanations on the same footing as the singular causes that compose them. The same factors that drive the attractiveness (or lack thereof) of singular-cause explanations drive that of plural causes.

The idea that causal cognition admits causes featuring several variables is not in itself new. In causal inference, researchers have studied how people infer conjunctive causes, that is factors that act in concert to produce an effect (Novick and Cheng 2004). The notion of a multivariate cause also plays a role in some theories of actual causation (e.g. Halpern 2015), and, in a different way, in philosophers' and economists' concept of *collective responsibility* (e.g. Arendt 1987; Miller 2001). To our knowledge, however, the literature on causal selection judgments has yet to engage with the notion of plural causes.

We present the first systematic study of plural causes in the context of causal selection.[1] This study has three objectives. First, we want to empirically establish the psychological reality and non-triviality of plural causes. We show that people's judgments about plural causes are sensitive to the prior probabilities of events, a key signature of causal-selection judgments. More importantly, we rule out a possible deflationary explanation for plural causes' sensitivity to probabilities: that subjects might formulate a judgment about a plural cause like $A \wedge B$ simply by combining in some direct way their judgment about the importance of the individual events A and B that compose it. In so doing we provide evidence that people treat plurals as full-fledged candidates for causal selection and engage with them in a holistic fashion.

Second, we want to show how considering plural causes can expand our understanding of the role of counterfactual reasoning in causal judgments. We show that models of causal selection based on the notion of counterfactual dependence can straightforwardly be extended to make non-trivial predictions about plural causes consistent with our findings. Counterfactual models consider that the causal impact of an event A on an outcome E is a function of the extent to which E depends on A across counterfactual worlds sampled in a certain way. We show that, similarly, people's intuitions as to the causal impact of a plural event $A \wedge B$ is to a large extent captured by the extent to which E counterfactually depends on it. At the same time, we highlight some ways in which other factors might contribute to the attractiveness of a plural explanation, above and beyond its mere counterfactual-dependence profile. These suggest ways in which current theories of causal selection could be improved to accommodate plural causes.

Third, we develop a perspective on the nature of the *representations* involved in subjects' causal selection judgments. Specifically, we propose that humans represent multivariate causes in causal reasoning in a way entirely analogous to how they represent the meanings of pluralities in natural language. The study of natural language reveals that plural entities such as can be expressed by the English noun phrases "Ann and Bill" or "the boys" are not represented as simple conjunctions of the atomic entities that constitute them. Instead, they possess non-classical semantic properties, in particular with respect to how they interact with negation. We provide evidence in the present study that causal judgments for multivariate events are best explained when we suppose that such events are treated as plural entities, with the same mathematical properties as the plural representations that underly our faculty for language. We conclude that theories from formal natural-language semantics offer great promise as mathematically rigorous theories of mental representations in the general-purpose language of thought.

3.1.1 Causation and causal selection

Humans are adept at representing the world through a web of causal relations between events. Representing causal relations allows people to make sense of what they observe, make predictions about what's to come,

[1] Our Experiment 1 was presented at the Forty-fifth Annual Meeting of the Cognitive Science Society and published in the society's non-archival proceedings (Konuk et al. 2023a).

and influence the future in some cases (Chater and Oaksford 2013; Gerstenberg and Tenenbaum 2017; Pearl and Mackenzie 2018; Sloman and Lagnado 2015).

In the psychological literature, people's causal knowledge is usually modeled through formalisms such as Causal Bayes' Nets or Structural Causal Models. These systems represent aspects of the world with variables, causal relations between these variables, and probability distributions (Pearl 2000). They appear as integral parts of accounts of psychological faculties and functions related to causation, such as causal inference and counterfactual reasoning.

One such causation-related function is causal selection: faced with a complex causal structure, humans will gladly *select* one cause (or, as we will show, more than one) as being more important than others. Moreover, they will assign different scores to different causal variables depending on how they perceive each of those variables as being *the* driver of the observed outcome.

Knowledge of the causal rule in the relevant system is of course one of the main factors determining the explanation humans will favor in causal-selection judgments. The other main driver of causal selection is the *normality* attached to events, a notion that combines the extent to which an event abides by moral or conventional rules, and the extent to which it was expected to happen, before it did happen (Icard, Kominsky, and Knobe 2017; Morris et al. 2019; Quillien and Lucas 2023).

The relationship between the causal rule that entangles events with the outcome, their normality, and causal-selection judgment can be complex. In a situation where several different variables are each individually *necessary* for an outcome, people tend to think of the *least expected* variables (the lightning bolt) as *the cause*, and comparatively disregard the importance of the most expected variables (the presence of oxygen), a pattern of judgment known as *abnormal inflation*. The converse tendency is observed in situations where all of the variables considered are each individually *sufficient* for the outcome to occur. In this case, people tend instead to think of the most normal events as the most important causes of the outcome (Icard, Kominsky, and Knobe 2017).

3.1.2 Defining the candidates for causal selection

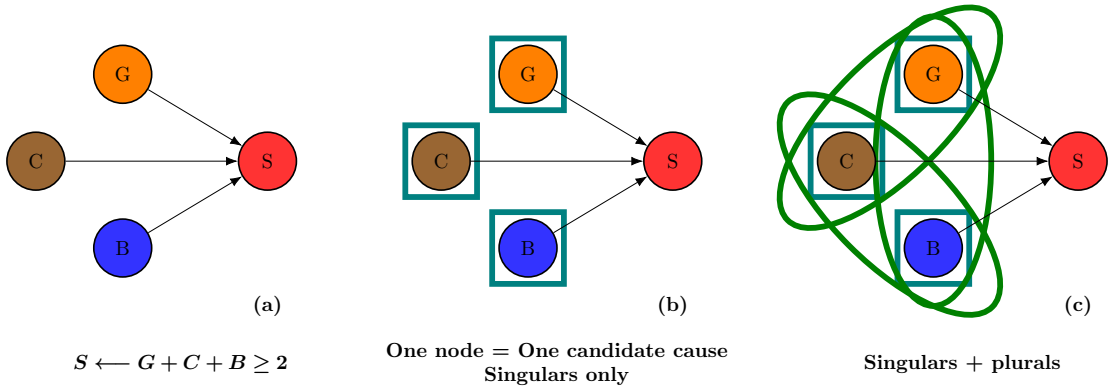
Causal selection is determined by an amalgam of the system's underlying causal rule and the normality of events. Thus, a standard procedure for formulating theories about participants' causal-selection judgments starts by building a causal model that formalizes their causal knowledge of the system.

Suppose for example that I get a stomachache shortly after having eaten a piece of Gouda cheese and a plate of pudding containing chocolate cake and blueberry pie. A causal model of this situation would feature one variable for each of the causes of my stomachache (i.e. one variable each for "eating the Gouda cheese," "eating the chocolate cake," and "eating the blueberry pie") as well as a variable for the effect ("having a stomachache"). The model also specifies a functional relationship between the variables, for example representing the fact that one develops stomach issues after eating too much, as schematized in Figure 3.1a.

As illustrated by this representational format, it is natural to think of the candidates for causal selection as particular realizations of the individual variables. If an individual equipped with the causal knowledge encapsulated by the model in Figure 3.1a wonders what *the cause* of their stomachache was, it may seem like they have to make a choice between the variables G, C, and B. This would directly identify the candidates for causal selection as the individual moving parts of the causal model, as represented in Figure 3.1b.

A striking feature of the psychological literature on causal selection is indeed that causal selection judgments are only ever queried at the level of singular variables (Kominsky et al. 2015; Morris et al. 2019; Quillien and Barlev 2022; Quillien and Lucas 2023; Sytsma 2020). Kinney and Lombrozo (2024) deserve an honorable mention in this connection however, since they compared participants' preferences for causal

Figure 3.1: A causal model for the relations between various dishes and my stomachache. **(a)** I develop a stomachache if and only if I eat two kinds of pudding or more. **(b)** The standard implicit assumption in the literature is that only single variables are candidates for causal selection. **(c)** We propose instead that causal judgment can also target plurals, for example pairs of variables.



generics ("X causes Y") mentioning one vs several variables. But their work was on type causation, while here we discuss token (actual) causation.

Concretely, when experimental participants are presented with a situation where an outcome depends on three different events A, B, or C, they are never asked to what extent a *plural* event like $A \wedge B$ can be considered the cause of the outcome. Intuitively though, causal explanations that mention combinations of variables can also be appealing. In our example above, saying that I got a stomachache "because I ate the entire dessert plate" might appear to be a better explanation than either "because I ate the chocolate cake" or "because I ate the blueberry pie" each on its own.

Note that allowing for many variables to feature in causal explanations does not eliminate the need for causal selection: one might want to mention several causes of an event without mentioning *all* of them. For example, one might think that "because I ate the blueberry pie" is a better explanation for my stomachache than "because I ate the entire dessert plate" if for example I eat chocolate cake at every meal, but add a blueberry pie on top of it only exceptionally. Ultimately, the best candidates for causal selection are those causes that participants see as most *crucial* in bringing about the outcome, whether these be singular or plural, and in principle we can only know what the best causal explanations are after considering the entire set of possible candidates, including plural causes, as illustrated in Figure 3.1c.

3.1.3 Counterfactual theories

To properly argue the point above, we first need to spell out what it means for a cause to be of a more or less crucial importance in bringing about an outcome. The notion we will rely on throughout this paper is rooted in counterfactual theories of causal selection (Icard, Kominsky, and Knobe 2017; Quillien and Lucas 2023).

Counterfactual theories of causal cognition in general build on the premise that humans represent causal relations between variables in terms of counterfactual dependence (Gerstenberg and Tenenbaum 2017; Halpern and Pearl 2005; Krasich, O'Neill, and De Brigard 2024; Lewis 1973b; Woodward 2003, 2006). The

notion that "C caused E" is taken to be roughly equivalent to the notion that "had C not happened, E would not have happened either."

In the case of causal-selection judgments, this is enriched with the important idea that the counterfactual dependence between C and E is evaluated not just in the actual world but in other possible worlds as well. Of particular relevance to this evaluation will be the possible worlds that are most *normal*, or *closest to the actual world* in which we are to select a cause (Lewis 1973b). Evaluating counterfactual dependence in these worlds is what allows a causal-selection judgment to provide explanations that are not just relevant to the situation under consideration, but also generalizable to other contexts (Hitchcock 2012; Lombrozo 2010).

We will limit our discussion in this article to two counterfactual theories (and accompanying models) of causal-selection judgment that (1) have been stated in full mathematical rigor and (2) have been submitted to experimental scrutiny, the Necessity and Sufficiency Model (Icard, Kominsky, and Knobe 2017, NSM) and the Counterfactual Effect-Size Model (Quillien and Lucas 2023, CESM). We chose to focus on these two theories because of their good track record in predicting participants' causal-selection judgments across a wide variety of tasks (Gerstenberg and Icard 2020; Gill et al. 2022; Henne et al. 2021, 2019; Kirfel and Lagnado 2021; Morris et al. 2019; O'Neill et al. 2022; Quillien and Barlev 2022).

The two theories see causal selection as a two-step process, the first of which identical, the second divergent. The procedure is as follows.

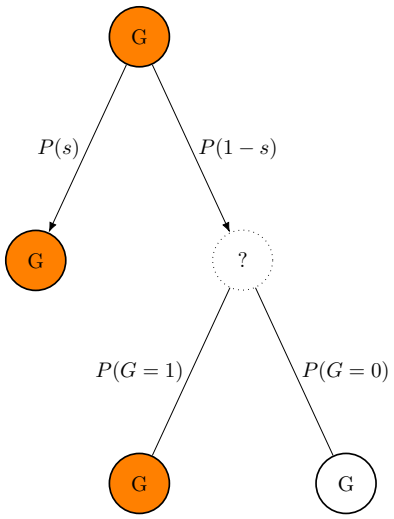
First, randomly sample a large number of counterfactual worlds. The sampling process operates at the level of the individual exogenous variables of the relevant causal model, that is the variables that have no parent in the causal graph. Across worlds, each of these variables is sampled with a frequency that is a function of two elements:

1. Its value in the actual world. Given a causal model with a set of exogenous variables E, and a valuation function $[\![\cdot]\!]^w$ that maps each variable in E to the value it has in a world of evaluation w, we can define a special world constant $w_@$ designating the actual world, that is the set of circumstances that in fact took place. The model includes a stability parameter s taking a value between 0 and 1, such that in every counterfactual world $w_i \in \{w_1, \ldots, w_n\}$ that it samples, each variable in E will have in w_i the same value that it has in the actual world $w_@$, with probability s (see Lucas and Kemp 2015; Quillien et al. 2023). This parameter is not present in the original version of the Necessity and Sufficiency Model, having been introduced in models of causal selection by Quillien and Lucas (2023). It can however be straightforwardly introduced into it, as we will do in this article.

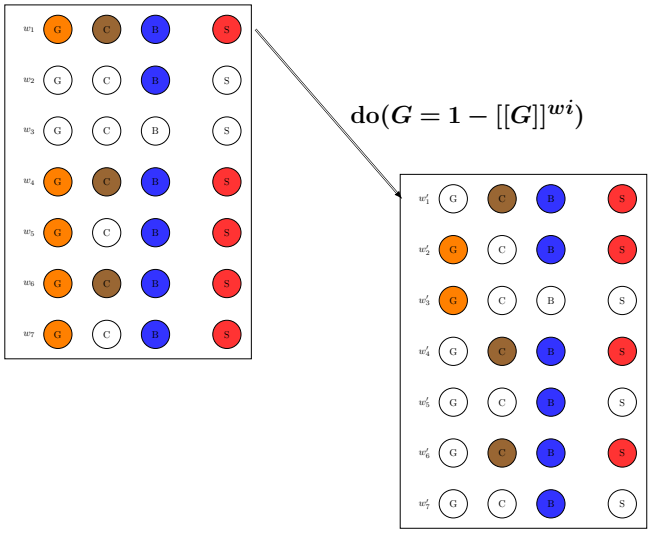
2. The variable's prior probability of occurrence. When a variable's value is not directly mapped to its actual world value, as will happen with probability $1 - s$, it is resampled from its prior probability distribution. This is where an event's sampling propensity (and from there, its causal score) gets to be sensitive to the normality attached to that event.

Consider the example of the causal system presented above relating variables G, C, and B to my stomachache. In the actual world, the variable G that encodes my eating Gouda cheese has value 1, meaning that event actually took place. As a result, when I sample counterfactuals to the actual world (as in Figure 3.2), I will with a probability s automatically represent myself eating cheese also in each of these worlds, as depicted in the left side of the tree in Figure 3.2a. In the worlds where I don't do that (the right side of the decision tree), the variable will be resampled from the prior probability on that event. Suppose for example that my eating cheese is a rather exceptional occurrence, such that $P(G) = 0.1$. In this case, I am much more likely to travel along the rightmost sub-branch of the tree in Figure 3.2a, and simulate worlds in which I didn't eat Gouda cheese than if I were accustomed to the fact and ascribed a higher prior probability to the event.

Figure 3.2: Sampling counterfactual worlds.



(a) Decision tree representing how the parameter s anchors the sampling of a variable to its real-world value (following Lucas and Kemp 2015).



(b) Representation of a sample of seven worlds; left: the same worlds, after an intervention flipping the value of G, with the resulting consequence on the outcome S.

All in all, the sampling propensity (Icard 2015) of any given exogenous variable V can be reconstructed as a function of the stability parameter s and that variable's prior probability $P(V)$, following the equation below, in the special case of interest here (binary variables, registering whether an event happens or not).

$$SP(V) = s \cdot [\![V]\!]^{w_@} + (1-s) \cdot P(V)$$

Once the exogenous variables of the system have been sampled in this way, one can simulate the outcome, which follows from the variables via the causal rule underlying this particular system, as in Figure 3.2b.

Second, compute the causal impact of a given variable V across those counterfactual worlds. The precise way to measure this impact is different in the two theories under consideration. In the NSM, causal impact is scored as the weighted sum of the following two factors.

1. A Necessity score: In each world w in which $[\![V]\!]^w \neq [\![V]\!]^{w_@}$, sample the outcome O from a probability distribution $P^v(O)$, where the value of each variable $V_j \in E$ other than V is switched to its actual world value $[\![V_j]\!]^{w_@}$. Then, count one point for the necessity score if the value of the outcome in the resulting world is different from the one that it had in the actual world (i.e. $P^v(O) \neq [\![O]\!]^{w_@}$), and zero points otherwise.

2. A Sufficiency score: For every world w in which $[\![V]\!]^w = [\![V]\!]^{w_@}$, sample the outcome from the probability of Sufficiency $P^{\sigma}_{V=[\![V]\!]^{w_@}}(O)$ of the event $V = [\![V]\!]^{w_@}$ for the outcome O. There is more than one way to define $P^{\sigma}_{[\![V]\!]^w=[\![V]\!]^{w_@}}$, but they make extremely similar predictions. The definition that turned out to have the best fit with the data from our experiments is the following.

$$P^{\sigma}_{V=[\![V]\!]^{w_@}}(O) = SP(O \mid do(V = [\![V]\!]^{w_@}), \neg[\![V]\!]^{w_@}, \neg O)$$

   From here, count one point for the sufficiency score if the value of the outcome thus sampled in w is the same as the value of the outcome in the actual world, and zero points otherwise.

Then, divide the number of points scored this way by the total count of worlds sampled. The dynamics of necessity and sufficiency scoring make it such that the necessity score is all the more important when the prior probability of an event is low, making it more likely to switch value across counterfactuals, whereas the sufficiency score is all the more important when this prior probability is high.

In the CESM, causal impact is computed using the same process in every world w_i, as follows.

1. Switch the value v of the variable V to a new, randomly sampled value v'. Then reevaluate the outcome in the new world w_i' where the value was switched. A representation of this resampling process is given in Figure 3.2b. The impact $K(V \rightarrow O_i, w_i)$ of V in the world w_i is then evaluated as

$$K(V \rightarrow O_i, w_i) = \frac{\Delta O}{\Delta V} = \frac{[\![O]\!]^{w_i} - [\![O]\!]^{w_i'}}{[\![V]\!]^{w_i} - [\![V]\!]^{w_i'}}.$$

   This equation can be glossed as follows. Whenever the outcome is switched in the same direction as the target variable (both from 1 to 0, or both from 0 to 1), V scores a point; when the outcome is unaffected, it scores none. When it moves in the opposite direction, it scores a negative point.

2. The causal impact is then normalized by the ratio of the standard deviations $\frac{\sigma_O}{\sigma_V}$, and averaged across worlds to get the causal impact score $K(V \rightarrow O_i, 0, [\![\cdot]\!]^{w_@})$ of the target variable for the target outcome

in the actual world. In simple conditions like the ones we will deal with in this article, the causal impact of V can be equated to the correlation coefficient between V and O across counterfactuals sampled at the first step.

3.1.4 Plurals in natural language

We hypothesize that humans represent and manipulate multivariate causes in causal selection in a way analogous to how they represent and manipulate the interpretations of plural noun phrases in natural language. Consider the following sentences.

(1) Who lifted this piano?
   a. It was Ann.
   b. It was Ann and Bill.

In response to a question as in (1), the answer in (1b) is just as natural as that in (1a): both sentences contain a perfectly coherent entity identified as the piano lifter, in (1a) a singular individual, in (1b) a plurality, or a collective. Accordingly, theories of plurality from linguistic semantics *generalize to the worst case*, taking the singular as the special case of the plural when cardinality is one, allowing for a unified account of singular and plural predication (Link 1983). Notice also that, in a situation where (1b) is the complete answer, a speaker might still accept the (partial) truth of a sentence like (1a), in virtue of the fact that Ann participated in the lifting event, especially if Ann's role was particularly important.

Analogously, in a causal system we can answer questions like "What caused my stomachache?" by pointing to individual variables ("It was the chocolate cake") or to multiple variables in one go ("It was the chocolate cake and the blueberry pie"). Just as in the language example, the null hypothesis is to expect the exact same mechanism to handle the singular and the plural case, since both the singular and the plural are perfectly coherent entities of the same type: causal entities which can be the target of causal selection.

And again just like the language case, in a situation where the multivariate cause "chocolate cake and blueberry pie" is the real culprit, one may still be inclined to (partially) accept the singular cause "chocolate cake," on the grounds that the chocolate cake participated in engendering my stomachache, especially if the role of the chocolate cake in bringing about this outcome was particularly important. We submit that this is why, although causal selection is at its heart sensitive to plural causes, experimental methods that haven't countenanced this possibly have still been very informative, just like judgments involving natural-language singular entities can be coherent and systematic and therefore informative, despite the fact that the fuller story involves plural entities.

Now, at first glance, in language in general as in causation, one would be tempted to propose that a plural entity should be understood as the mere *conjunction* of its individual constituents. That is, in the causal case, to claim that the plural variable "chocolate cake and blueberry pie" is to blame for my stomachache is actually mere shorthand for saying that the singular variable "chocolate cake" is to blame for my stomachache *and* the singular variable "blueberry pie" is to blame for my stomachache. The linguistic example in (1) already suggests that this might be missing something: if in actual fact it was Ann and Bill *together* that lifted the piano, then the plural in (1b) is the only precise way to describe the situation, a conjunction of the shape "Ann lifted this piano and Bill lifted this piano" would be quite inaccurate. Using language then as a source of conceptual possibilities with testable consequences, we might expect that plural causes should also display *togetherness* effects of this kind. The behavioral experiments we report on shortly will address this possibility head-on.

But plurals in natural language differ even more sharply from mere conjunction than by displaying this

kind of holistic effect. Of special relevance to what's to come is the fact that the negation of a plural does *not* correspond to the negation of a conjunction.

(2) Ann and Bill don't speak German.
   a. Either Ann doesn't speak German, or Bill doesn't, or neither does.
   b. Neither Ann nor Bill speaks German.

The most natural reading of (2) is as in (2b), and not as in (2a) as the standard Boolean semantics for "and" and "not" might lead us to expect (Krifka 1996; Lappin 1989; Löbner 2000; Szabolcsi and Haddican 2004). Thus, if multivariate causes are *plurals* in any interesting sense, we must consider the possibility that they behave like linguistic plurals, suggesting perhaps a non-flat weighting of the three in-principle ways of falsifying a conjunction, strongly favoring the possibility where all variables constitutive of the plural are individually negated.

For the purposes of this article, this last possibility will be particularly relevant in situations where subjects are tasked with formulating judgments about *negative* outcomes. That is, for example, when they have to explain *losing* a round of a game, in a situation where they have only explicitly been instructed in the conditions for *winning*. For producing such judgments will require them to internally use some equivalent of negation on their mental representation of the winning conditions. Our Experiment 2 will address this question, among others.

Lastly, a word is warranted on the theories of plurality from linguistics which describe and partly explain these facts. Since Link's (1983) seminal work on the logic of plurality, the consensus view has been that plural entities require their own dedicated mathematical structures for representation, and cannot be subsumed as special cases of Boolean conjunction or Boolean disjunction. Specifically, plural entities are formed out of singular entities ("atoms" in the formal semantics jargon) by a *mereological sum* operation. Mathematically, this is a join operation, giving rise to a join semilattice. Mixing and matching elements of the algebraic and order-theoretic characterizations of join semilattices for the sake of clarity and brevity, these are structures $\mathfrak{A} = \langle A, \oplus \rangle$, where A is the set of singular and plural entities and $\oplus$ is the plural-forming (join) operator, such that for any $a, b \in A$, we will also find $a \oplus b \in A$, representing the smallest plural containing both a and b. $\mathfrak{A}$ also comes with a partial order $\leq$ over A which formalizes the notion of containment, so that, for example $a < a \oplus b$. From this point onward the details differ in various accounts, but all share this algebraic structure, crucially distinct from the algebra induced by the more familiar Boolean connectives. This structure provides the mathematical degrees of freedom required to now define both *cumulative* and *distributive* predication (Ann and Bill lifting a piano together vs. Ann lifting a piano and Bill lifting a piano), and to describe the observed *homogeneous* interaction with negation (Krifka 1996; Križ and Spector 2021; Löbner 2000).

This is all the detail we can allow ourselves on this topic, short of reviewing an extensive and technically involved literature from formal semantics. The readers whose curiosity we've managed to whet will find a concise but mathematically rigorous introduction to these tools and their applications in an excellent handbook chapter by Champollion and Krifka (2016). Our goal with this brief illustration of the algebraic structures universally used in plural semantics is only to give a taste of what it might possibly mean to define a non-classical, non-Boolean, conjunction-like operation for plurality which is nevertheless entirely mathematically rigorous and intelligible. We will address why and how this kind of mathematical clarity about natural-language meaning matters in psychology in our general discussion.

3.1.5 Extending causal selection to plurals

The idea that people represent plural events produces testable predictions. For starters, an alternative view on such plural causal judgments could contend that people only ever have direct intuitions about the causal responsibility of the *individual* variables in their causal model. People might make judgments about a plural cause say by adding up or averaging the individual causal strengths of its constituent variables. For example, to compute how much they agree that "eating the chocolate cake and the blueberry pie caused the stomachache," people might first compute their agreement with "eating the chocolate cake caused the stomachache," "eating the blueberry pie caused the stomachache," and so on. Then they might somehow aggregate the causal strength of each individual variable. We will call this the *linear combination* hypothesis: a systematic simple combination of the scores for singular variables might underwrite participants' judgments about plural causes. This hypothesis is deflationary with respect to the psychological reality of plural causes in that it holds that people can make plural-cause judgments when prompted to do so, but they cobble them together from more primitive representations of causal strength at the level of individual variables. This means that the cognitive process underlying causal selection is ultimately still only ever deployed at the level of singular causes. In keeping with our language analogy, this would mean that the holistic, *togetherness* effects we find with linguistic plurals do not exist, or are trivial, in the domain of plural causes.

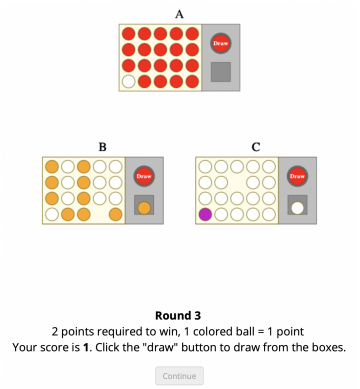
In contrast, we consider the possibility that plural causal judgments are the output of a *holistic* computation. Under this possibility, the same cognitive process that allows people to formulate causal judgments about singular variables is deployed at the level of combinations of variables, yielding quantities that can on occasion diverge significantly and non-linearly from the causal judgments for constituent variables. We consider in detail how such a hypothesis could be implemented in models of causal selection in the next section. But the general idea can be explained rather simply: to assess the causal importance of a plural event $A \wedge B$ is to look at the causal impact that this event has on the outcome of interest, using the same measures of causal impact that were detailed above for singular variables. This holistic computation will sometimes lead to different predictions than the hypothesis that consists in simply computing the impact of A, of B, and then combining them. For example, for a plural event $A \wedge B$ to be the most highly rated among possible candidate causes, the singular events A and B need not necessarily be the most highly rated among singular causes. In order to tease apart these two hypotheses in the case of causal selection, we need to identify contexts where they make different predictions about causal-selection judgments. This is what we do in the two experiments presented in this paper.

3.2 Experiment 1

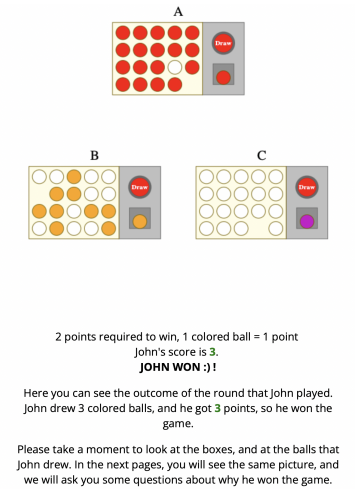
Our first experiment has the following goals.[2] First, if plural causes are processed as genuine causes by the mind, factors that are known to affect causal-selection judgments should influence judgments about a plural cause. In particular, the probability of an event is known to affect judgments about whether that event caused an outcome (Morris et al. 2019). We expect analogous patterns for plurals: varying the probability of events should affect causal judgments about whether a conjunction of these events caused the outcome. Second, we aim to rule out a deflationary *linear combination* account of the impact of probability on participants' judgments. Evidence of non-linearity in causal judgments would constitute stronger evidence for the psychological reality of plural causes in human causal selection. We design a situation where both the CESM and the NSM predict that the causal strength of plural variables will not be a linear combination of the

[2]This study was reported at the Forty-fifth Annual Meeting of the Cognitive Science Society and published in the society's non-archival conference proceedings (Konuk et al. 2023a). Our writing in this section borrows directly from this preliminary report.

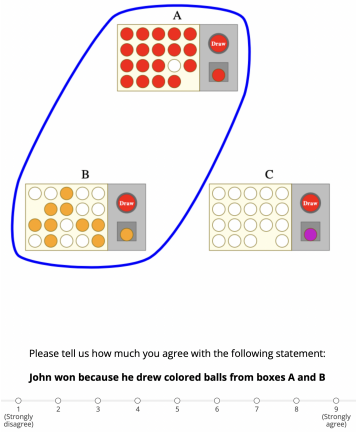
Figure 3.3: The three phases of Experiment 1



(a) First, participants played 10 rounds of the game to get familiar with the rule and the underlying probabilities.



(b) They then observed the outcome of a round of the game played by a different player named "John."



(c) We queried their judgments, supplementing the plural that occurred in the sentence in boldface with visual highlighting of the queried variable(s).

score of individual variables. We compare their predictions to those of a null model that tries to predict the score of plural causes as a linear combination of the scores of individual variables.

3.2.1 Methods

Design and materials

We adapted a paradigm developed by Quillien and Lucas (2023). Participants made judgments about a game of chance, in which one randomly draws balls from a set of urns, and wins by getting enough colored balls (see Figure 3.3 for illustrations). Participants observed a fictitious player draw a colored ball from each of three urns (labelled A, B, and C) and win the game as a result. Then they were asked to make a causal judgment about each singular cause (e.g. whether getting a colored ball from urn A caused the player to win the game), and about each pair of causes (e.g. whether getting a colored ball from urns A and B caused the player to win the game). For exploratory purposes, we also asked participants to make a causal judgment about the triplet (getting a colored ball from A, B, and C). We manipulated the prior probability of each outcome within participants by varying the proportion of colored balls in each urn, with probabilities of 0.05, 0.5, and 0.95 (Figure 3.3). We will refer to the three different urns as the LOW, INTERMEDIATE, and HIGH urns, respectively. The rule of the game, which was directly revealed to the participant at the outset, was that the player wins if they get two colored balls or more. This corresponds to the causal model below.

$$\text{WIN} := A + B + C \geq 2$$

Predictions

This paradigm provides a context where the linear and the holistic extensions of the models we outlined above make clearly different predictions. The CESM predicts that participants' singular causal-strength estimates should follow a particular ranking: INTERMEDIATE $>$ LOW $>$ HIGH, for any value of the *s* parameter. This is because, across possible counterfactual alternatives to what happened, there is a high correlation between getting a colored ball from the intermediate probability urn and winning the game. These predictions partially match participants' judgments collected in the previous iteration of this paradigm run by Quillien and Lucas (2023), where judgments were collected for singular variables only, and in which participants' responses followed the ranking: INTERMEDIATE probability urn $>$ LOW $\approx$ HIGH (there was no significant difference between the ratings for the low probability and the high probability urn).

If one considers a linear extension of the CESM, where participants simply combine the causal strength of individual variables to make plural cause judgments, or simply a linear combination of participants own recorded judgment, they should consider that the pair LOW & INTERMEDIATE should have greater-than-or-equal causal strength to the pair HIGH & INTERMEDIATE, because the singular LOW has higher causal strength than HIGH.

In contrast, if participants judge the causal strength of plurals via a holistic computation, they should rate the pair INTERMEDIATE & HIGH as highest. For, across possible counterfactuals, there is a high correlation between getting a colored ball from these two urns and winning the game. Intuitively, since drawing a colored ball from the low-probability urn is rare, and given that at least two balls are needed to win, most worlds where the player wins the game will be worlds in which they do so by getting a colored ball from the INTERMEDIATE and HIGH urns. This prediction is true for any value of the *s* parameter in the holistic version of the CESM. It is also true for the holistic version of the NSM, although in that case it does reverse the ranking that the NSM expects for singulars.

Procedure

Participants first completed ten rounds of the game, presented with urns as in Figure 3.3a. We pseudo-randomized the draws in such a way as to get participants to internalize the probabilities associated with each urn and how they connected to the outcome. Then participants saw the outcome of a round of the game played by another (fictitious) player, who drew a colored ball from all three urns, thereby winning with 3 points, as in Figure 3.3b. They were asked to rate the causal strength of each individual draw, as well as that of every combination of two or three draws for the winning outcome, on a Likert scale from 1 to 9 (strongly disagree to strongly agree), as in Figure 3.3c. We used these likert scale ratings as people's ratings for a given cause has been shown to correlate strongly with how likely they are to single out that same variable as "the cause" of an outcome (Morris et al. 2019). For the singulars, participants were asked to rate their agreement with the statement "John won because he drew a colored ball from box [urn]." For the plurals, they rated their agreement with "John won because he drew colored balls from boxes [urn1] and [urn2]." Each question was displayed on a separate page, next to the urns that displayed the outcome of the fictitious player's draw. The letters indexing the urns, as well as the colors of the balls, were randomized across participants but were kept the same within a participant. Half of the participants were asked about the singulars first, and then about the pairs. The other half were asked about the pairs first, and then about the singulars. All participants were asked about the triplet at the very end. Within one class of questions (singulars vs. plurals) we randomized the order of presentation of questions. Finally, participants completed a brief demographic questionnaire and were redirected to Prolific for payment. We coded the experiment in the `jsPsych` library (De Leeuw 2015), with custom plugins for displaying urns developed in our lab.

Table 3.1: ANOVA for singular causal-selection judgments, predicting urn ratings from urn probabilities and urn order of presentation.

Factors	Mean Sq	F-score	p-value	η_p^2
Probabilities of the urns	32.984	4.492	< 0.012	0.008
Order of presentation	138.808	18.904	< 0.0001	0.020
Probabilities:order	3.177	0.433	> 0.640	0.001

Participants

We recruited 400 participants from all English-speaking countries from Prolific. This sample size was inspired by the one used by Quillien and Lucas (2023), who used a comparable sample size (290 participants), for a study with similar design. We excluded from subsequent analysis 44 participants who failed to answer either of two elementary comprehension questions that checked their understanding of the rules of the game, leaving a total of 356 participants for analysis.

Transparency and openness

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. All data, analysis code, and research materials are available at `https://osf.io/43m5d/?view_only=3d26a80b8e394fa9ad7792a690de8fe6`. Data were analyzed using `R` (R Core Team 2013), version 4.3.3 and the package `ggplot2`, version 3.5.0. All studies we report in this article received ethics approval by the *Comité d'évaluation de l'éthique de l'INSERM*, under research protocol *Le langage et les capacités cognitives connexes*. All studies were conducted entirely in English. We did not preregister the studies.

3.2.2 Results

We first report analyses using standard statistical tests. Then we report the fit of computational models of causal judgment.

Basic results patterns

Prior probability affects both singular and plural causal judgments. Results are plotted in Figure 3.4. We ran two two-factor repeated-measure ANOVAs, one for each main type of cause queried (*singulars* and *pairs*), using urn probabilities and order of presentation as predictor variables, and participants' responses as the dependent variable. Results are in Tables 3.1 and 3.2. There was a main effect of prior probability on participants' causal judgments, for singular as well as for plural causes ($p < 0.020$ in both cases), consistent with our expectation that participants' judgments for plural causes should be sensitive to probabilities just like for other actual cause judgments.

The order of presentation/querying singular and plural selection judgments also had a significant effect ($p < 0.001$) on the ratings for singulars: singular causal judgments were lower when presented after the plurals. There was however no interaction effect between urn probability and order of presentation, suggesting that the impact of probability on causal estimates did not vary depending on the order in which questions were asked. Therefore we drop this variable (order of presentation) from later analyses.

Figure 3.4: Mean ratings by question type, along with predictions for each theory under consideration. The error bars represent the standard error of the mean. The *linear combination* theory (purple) predicts that the score of the LOW & INTERMEDIATE and INTERMEDIATE & HIGH pairs should be equivalent, when in fact we see a significant difference between them, in accordance with the *holistic* versions of the two counterfactual models: the Counterfactual Effect Size Model (Quillien and Lucas 2023) (in Red on the plot) and the Necessity and Sufficiency Model (Icard, Kominsky, and Knobe 2017) (in Blue).

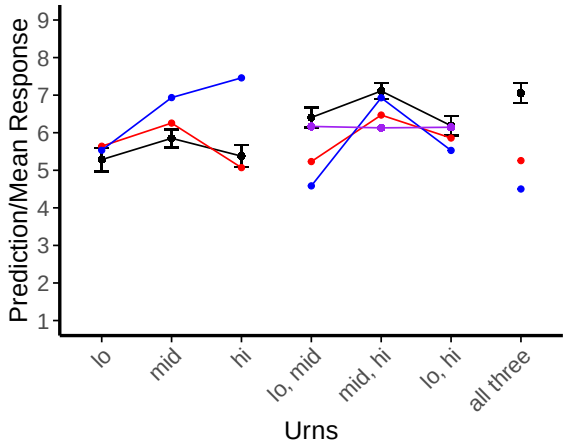


Table 3.2: ANOVA for pair causal-selection judgments.

Factors	Mean Sq	F-score	p-value	η_p^2
Probabilities of the urns	83.018	14.646	< 0.000001	0.030
Order of presentation	21.747	3.837	> 0.050	0.004
Probabilities:order	2.303	0.406	> 0.660	0.001

The causal strength of plural causes is not a linear combination of the causal strength of individual variables. The pattern of responses for singular variables replicated the patterns obtained by Quillien and Lucas (2023). Judgments for the INTERMEDIATE urn were higher than judgments for the LOW urn, $t(315.41) = -2.70$, $p < 0.008$, and the HIGH urn, $t(325.85) = -2.08$, $p = 0.038$. The difference between the LOW and HIGH urns was not significant, $t(350.59) = -0.63$, $p > 0.520$.

We can use these results to test the *linear combination* hypothesis, according to which participants derive their plural-cause strength estimates by adding up or averaging their estimates for the individual variables that compose a given plural. If this were correct, participants should give the same causal strength estimate for the two plural causes LOW & INTERMEDIATE and INTERMEDIATE & HIGH, since their estimates for the singular causes LOW and HIGH are not significantly different from each other. By contrast, both the holistic CESM and the holistic NSM predict a sharp difference between these two kinds of plurals, with the plural cause INTERMEDIATE & HIGH being rated higher (Figure 3.4).

Consistent with the holistic CESM, judgments about the INTERMEDIATE & HIGH pair were higher than for the LOW & INTERMEDIATE pair, $t(355) = -4.67$, $p < 0.001$, and higher than for the LOW & HIGH pair, $t(355) = 6.858$, $p < 0.001$. In slight deviation from the CESM and NSM's predictions however, judgments

Table 3.3: Results of the ANOVA: estimate for pairs ∼ est. for singular-1 × est. for singular-2.

Factors	Mean Sq	F-value	p-value	η_p^2
LOW	100.692	17.743	<0.001	0.016
INTERMEDIATE	20.795	3.664	0.056	0.003
HIGH	15.934	2.808	0.094	0.003
LOW:INTERMEDIATE	6.636	1.169	0.280	0.001
LOW:HIGH	0.414	0.073	0.787	0.000
INTERMEDIATE:HIGH	46.642	8.219	0.004	0.008

Table 3.4: Comparison between two models: the linear combination model of plurals (means of singulars + intercept), and the means of singulars + question model.

Models	LogLik	Df	χ^2	p-value	BIC
Means sing	-891.89	3			1804.709
+ Question	-878.13	5	27.53	< 0.00001	1791.126

for the *low, intermediate* pair were higher than for the *lo, high* pair, $t(355) = 2.369$, $p < 0.025$ (Figure 3.4).

We conducted two more analyses to rule out the linear combination model. First, we ran a one-way repeated-measure ANOVA, predicting judgments for the pairs (LOW & INTERMEDIATE, INTERMEDIATE & HIGH, and LOW & HIGH) from judgments for the singulars (LOW, INTERMEDIATE, and HIGH), as well as their interactions, as within-participant factors. Each plural pair was regressed only on the values of the two singulars that comprised it.

The linear combination theory predicts that there should be no significant interaction: a participant's causal judgment for a given *singular* variable should have the same impact on every plural cause in which it features. One's estimate for the singular INTERMEDIATE, for example, should have an equal impact on one's estimate for INTERMEDIATE & HIGH and for LOW & INTERMEDIATE.

We find evidence against the linear combination theory (Table 3.3). There was a significant interaction between the INTERMEDIATE and HIGH urns, $p < 0.005$. In addition, the main effects of the singular judgments were not significant, for all but the LOW urn.

Second, we fitted linear multilevel regression models on participants' responses for pairs. Specifically, we compared the predictive performance of two different models on participants' plural-cause estimates. The first one used as predictor the average of the two singular-cause estimates for the variables contained in a given plural (computed on a per-participant basis), plus a random intercept. The second model also included the question asked (that is, the specific plural being queried) as predictor. A likelihood ratio test shows that adding question as a predictor significantly improves the fit of the model, $\chi^2(5) = 27.53$, $p < 0.001$ (Table 3.4). Again, this is inconsistent with the linear combination account.

Computational modeling

We computed the predictions of two recent counterfactual models of causal selection, the Counterfactual Effect Size Model (Quillien and Lucas 2023) and the Necessity and Sufficiency Model (Icard, Kominsky, and Knobe 2017), presented in the introduction. Our implementation follows the one given in Quillien and Lucas (2023).

Table 3.5: Table of model comparison, Study 1, excluding the triple. The AIC and BIC values are computed for mixed effects models, including group and a random effect for participants.

Model	AIC	BIC	Cor.
CESM	9929.31	9957.65	0.54
NSM	9962.06	9990.4	0.24
Considering only the pairs			
CESM	4692.11	4716.978	0.83
NSM	4674.355	4699.222	0.80
Empirical average	4692.942	4717.81	-0.65

For each question we report on below, we generated causal judgments for the CESM using a process of counterfactual sampling. We generated predictions for the CESM by simulating 10^5 possible rounds of the game according to the rule, what was the case in the situation described to participants, and the sampling model described by the CESM. We computed CESM judgments for an event as the correlation between that event (for instance, whether the player draws a colored ball from urn *A*) and the outcome of the game (whether the player wins the game), across simulations. We computed NSM judgments analytically, as the sum of the variables' sufficiency and necessity scores across worlds.

We fit the value of the stability parameter s for both models by finding the value of s that results in the best fit between model judgments and average participant judgments across all seven questions. We quantified model fit by looking at the likelihood of mean answers per question under a normal distribution centered on the model's predictions, with a standard deviation fitted across questions.

We identified the best fit value via a grid search, exploring a wide range of values for the parameter s, crossed with different values for a scaling parameter γ (applied to a model's predictions as an exponent $prediction^{\gamma}$. The point of γ was to avoid situations where one model would systematically overshoot or undershoot actual participant answers, as the models are not meant to predict the exact value of participants' judgments, but only the relative difference between one variable and the next. Our technique here was analogous to that of Griffiths and Tenenbaum (2005).

For the CESM, the best fitting value was $s = 0.89$, with $\gamma = 0.26$. For the NSM, the best fitting value was $s = 0.71$, with $\gamma = 2.93$.

In our implementation, to assess the causal strength of plural causes a model assumes that people compute the causal strength of the conjunction of all variables contained within that plural. For instance, the CESM computes the causal strength of LOW & HIGH by computing the correlation between the binary variable LOW $\wedge$ HIGH (which has value 1 if both LOW and HIGH have value 1, and 0 otherwise) and the outcome.

The predictions of the models are plotted in Figure 3.4. Table 3.5 details the comparison. Overall, the CESM's prediction had the best fit to human judgments in this experiment, although the NSM had the best fit when models were compared on pairs of variables only. We also compared the models' performance on the judgments for pairs to a null model that used as predictor for each pair the average of mean human judgments for each singular variable contained within a given pair, as plotted in Figure 3.4. Both counterfactual models proved significantly better than this linear predictor (Table 3.5).

3.2.3 Discussion

We find evidence that, when people make a judgment about whether events A and B caused an outcome, their judgments track the correlation between the conjunction of A and B and the outcome, across counterfactuals. Concretely, in our experiment, winning the game is in general strongly associated with getting a ball from both the intermediate- and high-probability urns, and people judged that combination of events to be highly causal. Importantly, this effect is inconsistent with a simpler account, according to which people's judgments about plurals are cobbled together from their causal intuitions about each individual variable in the plural.

Judgments about plural causes are affected by the prior probability of their constituent variables, but cannot be derived from the causal strength of these individual variables. As such, our results are in general consistent with the predictions of simple extensions of recent counterfactual models of causal selection (Icard, Kominsky, and Knobe 2017; Quillien 2020; Quillien and Lucas 2023), augmented with the assumption that people judge plural causes in a holistic manner.

At the same time, these findings raise new questions about the psychology of causation. Presently we highlight two of these questions, which we investigate in Experiment 2.

First, participants in this study found the plurals overall more appealing than the singulars, a tendency which the counterfactual models we considered did not capture. Participants might have felt that plurals provided more exhaustive descriptions of the event: they give more complete information about what happened, in addition to why it happened. We also find that this effect is accentuated when singulars are presented after plurals. Making judgments about plurals first might highlight to participants the descriptive incompleteness of singulars. This finding suggests an interesting tension between two potential desiderata of causal judgment: highlighting the variables that were most causally important to the outcome, and providing an exhaustive list of the causal factors. If so, this calls for an investigation into the relative importance of these two pressures in participants' causal-selection judgments. When, for example, adding a variable weakens the counterfactual dependence profile of the resulting plural, such as when the plural doesn't explain the outcome appreciably better than one of the singular variables within it, will participants still show a preference for plurals, on account of their greater completeness?

Second, a notable feature of this first experiment is that the causal structure used a simple additive rule (i.e. the player wins the game if their score is above a certain threshold). As such, there is a sense in which the variables each have an independent incremental causal effect on the outcome.

What will participants' plural-cause judgments look like in a causal structure where some conjunctions of events directly feature as such in the causal model that generates the outcome? Consider for example the causal rule $(A \wedge B) \vee C$. Here the urns A and B are specifically connected in the logical structure. Generalizing somewhat, our question here is: when an outcome specifically depends on the joint occurrence of A and B, should that make the plural cause $A \wedge B$ a more natural causal explanation than a potential alternative $A \wedge C$, even if C also makes an important contribution to the outcome?

3.3 Experiment 2

Experiment 1 established the psychological reality and relevance of plural causes for causal-selection judgments. Building on its findings, Experiment 2 expands our exploration of plural causes in four directions.

First, we provide additional evidence against deflationary interpretations of plural causes. We give more examples of situations in which people's plural-cause judgments cannot be straightforwardly derived from a linear combination of their singular-cause judgments, to confirm the results obtained in the first experiment.

Second, we explore whether there is a robust bias toward preferring causes that contain more variables. In

Experiment 1, participants gave overall stronger scores to plural causes than singular ones. Experiment 2 investigates whether this pattern always holds.

Third, we explore a richer causal structure. Here, two urns contain purple balls, and two urns contain orange balls. The player can win the game by getting either two purple balls or two orange balls, where "or" is meant inclusively. Formally, winning can be triggered by either of two distinct sufficient conditions $A \wedge B$ and $C \wedge D$, each a conjunction of two variables. This corresponds to the rule

$$\text{WIN} := (A \wedge B) \vee (C \wedge D) \tag{3.1}$$

Fourth, we explore participants' judgments in situations where they have to explain a *negative* outcome. In the context of our experiment, this amounts to explaining a loss. We call losing a negative outcome here in the sense that subjects have only been explicitly instructed in the conditions for winning. The conditions for losing are only implicit in these instructions, as the negation of a winning outcome. As we detail presently, this contrast opens interesting possibilities about the representations participants might have for losing conditions.

Finally, Experiment 2 was also designed to collect many more data points per participant, increasing our statistical power compared to Experiment 1. We ask each participant about the outcome of four possible rounds of the game (as opposed to just one outcome in Experiment 1), collecting a total of 36 causal judgments per participant.

3.3.1 Negative outcomes and plurals

The subject of responsibility attributions for negative outcomes is relatively understudied, and it is by no means a trivial question how such attributions will relate to the patterns found in judgments for positive outcomes.

Recently, in a study of *ex-ante* responsibility judgments (that is, judging the importance of various causes before any outcome has effectively occurred), Gerstenberg, Lagnado, and Zultan (2023) observe that subjects' estimates are better captured by a measure that tracks a variable's contribution to wins than one that tracks its contribution to losses (or some hybrid of the two). This suggests that positive outcomes are taken to be the explananda by default for responsibility judgments. Explaining negative outcomes would on the other hand demand some additional mental operations.

Yet one might expect that the operations in question should be trivial, especially for binary variables. Indeed any process that assigns responsibility for a win could in principle be straightforwardly repurposed for losses, simply by moving the target. Erstwhile "wins" now count as losses and "losses" become wins, as far as assigning credit to causes. Translated into counterfactual models, this would simply amount to tracking how different causes co-vary with the *classical logical negation* of the winning conditions across counterfactuals. But available evidence suggests a more complex picture. Gerstenberg and Icard (2020) looked at causal selection judgments in a billiard-ball setting involving simple conjunctive or disjunctive rules. They collected subjects' estimates in positive cases where two events A and B happened, and in negative cases where both A and B failed to happen. As they noted, judgments for negative outcomes in the *disjunctive* cases where $E := A \vee B$ were neatly captured by treating negative outcomes as the classical negation $\neg E := \neg A \wedge \neg B$. But the same strategy did not work in the conjunctive case where $E := A \wedge B$. As we will see shortly, we found related results in our Experiment 2, where subjects' judgments did not fit the pattern expected by the classical interpretation of negated conjunctions. Now, why might negation interact so differently with disjunction (classically) and with conjunction (non-classically)?

Plural entities in natural language have the logically surprising feature that negation applies homogeneously to each individual in the plurality. Consider the examples of plurals in (3) and (4), and the putative

interpretations for the negated plural (4) in (4a) and (4b).

(3) The boys did their homework.

(4) The boys didn't do their homework.
   a. None of the boys did his homework.
   b. At least one of the boys didn't do his homework.

Sentence (3) means that *every* boy did his homework, with some tolerance for exceptions which needn't concern us here (Križ and Spector 2021). Sentence (4) then ought to be simply the negation of (3), which would amount to the interpretation paraphrased in (4b). Yet, the negated plural in (4) has a much stronger interpretation, to the effect paraphrased in (4a). In general, negated plurals are interpreted in this unexpected way, from the standpoint of classical logic (Krifka 1996; Lappin 1989; Löbner 2000). This observation applies to plurals as in (4), generated by a noun phrase with plural morphology "the boys," but also to plurals formed by means of an explicit conjunction: a sentence like "John and Mary didn't countenance this hypothesis" means that neither John nor Mary considered this hypothesis, not merely that at least one of John or Mary failed to consider it (but see Szabolcsi and Haddican 2004, for evidence of cross-linguistic variation on the available interpretations).

In light of these observations, we hypothesize that participants in our experiment might track the losing conditions of the rule in equation 3.1, which is a disjunction of plural terms, in this non-standard way. This would amount to the stronger loss conditions at the end of equation 3.2 below, which we preface with '$\not\equiv$' to indicate that it violates classical-logical equivalence.

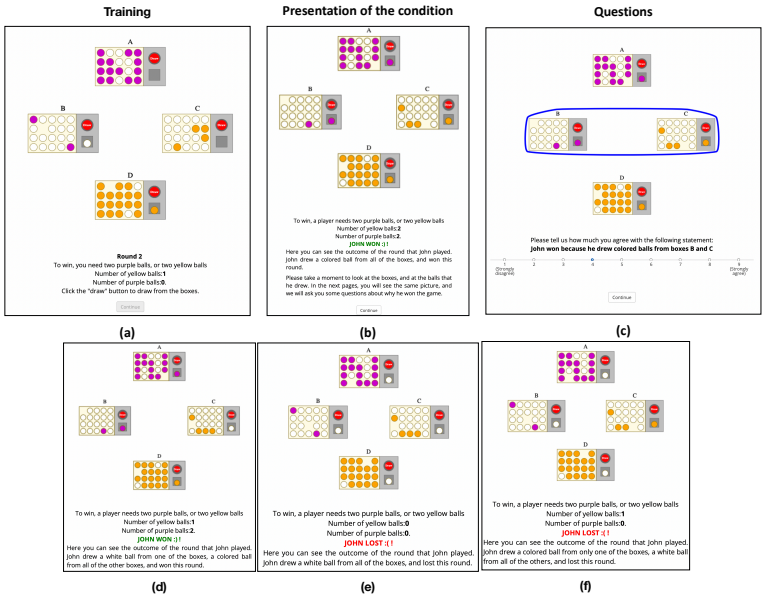
$$\begin{aligned} \text{LOSS} &:= \neg((A \wedge B) \vee (C \wedge D)) \\ &\equiv \neg(A \wedge B) \wedge \neg(C \wedge D) \\ &\not\equiv \neg A \wedge \neg B \wedge \neg C \wedge \neg D \end{aligned} \tag{3.2}$$

Such a reading of the losing conditions would yield very different judgments for the negated conditions than the reading that arises from the *classical logical negation* of the conditions for winning the game, represented in equation 3.1 above, as follows.

$$\begin{aligned} \text{LOSS} &:= \neg((A \wedge B) \vee (C \wedge D)) \\ &\equiv \neg(A \wedge B) \wedge \neg(C \wedge D) \\ &\equiv (\neg A \vee \neg B) \wedge (\neg C \vee \neg D) \\ &\equiv (\neg A \wedge \neg C) \vee (\neg A \wedge \neg D) \\ &\quad \vee (\neg B \wedge \neg C) \vee (\neg B \wedge \neg D) \end{aligned} \tag{3.3}$$

To be very clear, we do *not* expect that people will *misinterpret* the rules of the game. We do not expect that they would classify, say, a round of the game where the player only draws colored balls from urns A and C as anything other than a loss, any more than we expect English speakers that take (4a) to be the natural interpretation for (4) would mistake (3) for a true sentence, had exactly one boy done his homework. Instead, our hypothesis is that the tendency to negate plurals in this homogeneous way will affect the particular instances of losses that *come to mind implicitly* and, from there, the target that subjects will try to match as they compute how much each cause contributes to losing across counterfactuals. We spell out a model implementation of this idea in the section dedicated to losing rounds below.

Figure 3.5: Top: The three phases of Experiment 2, analogous to those of Experiment 1, by order of presentation to the participants. The condition presented here as example is the OVERDETERMINED NEGATIVE condition. Bottom: the three other conditions presented to participants: (d) TRIPLE-1 ; (e) OVERDETERMINED NEGATIVE ; (f) TRIPLE-0



We refrain for now from any discussion of the possible *reasons* and *mechanisms* whereby participants might engage in this language-like treatment of negated plural causes, at this point we mean only to point out that this is a plausible hypothesis worthy of testing. We will address the theory questions in the general discussion.

3.3.2 Methods

Design and materials

The methodology was similar to that of Experiment 1. We presented participants with a simple game of chance. This time, the game involved four urns, with two different colors, purple and yellow (Figure 3.5). We randomized the assignment of colors, but always in such a way that urns A and B were of one color, and urns C and D of the other color. To win a round of the game, one needed to draw "two purple balls or two yellow balls."

While we randomized the specific urns' indices and their spatial arrangement for each participant, for simplicity here we refer to a consistent arrangement as depicted in Figure 3.5, where urn A has 14 colored balls, urn B 2, urn C 4, and urn D 19 colored balls. These induce different prior probabilities of drawing a colored ball out of each urn, such that $P(A) = 0.7$, $P(B) = 0.1$, $P(C) = 0.2$, and $P(D) = 0.9$. Throughout the experiment, the urn containing 14 colored balls and the urn containing 2 colored balls were always of the same color, while the other two urns (19 and 4 colored balls) were of the other color, so that each color would contain one high probability and one low-probability urn.

Procedure and participants

As in Experiment 1, participants first had the opportunity to familiarize themselves with the game and the rule determining a winning outcome, as well as with the underlying probabilities, by playing the game for ten rounds, as in Figure 3.5a. Urn draws and outcomes at this stage were pseudo-randomized in such a way as to reflect the underlying probabilities.

After they played ten rounds of the game, they saw the outcomes of rounds played by another player named John (as in Figure 3.5b) and were asked to rate on a Likert scale from 1 to 9 the causality of certain events, both singular and plural. Specifically, we queried their causal judgments by asking them the extent to which they agreed (on a 1–9 scale) with a sentence that followed the template: "John won (/lost) because he drew colored (/white) balls from box(es) [XYZ]." Figure 3.5c shows an example.

All participants saw four different rounds of the game played by John, one at a time, and provided their judgments after each round. All the rounds were played with the same underlying rule and the same urns in the same display as the one participants had previously been familiarized with. Each trial differed only in the outcome of the draw made by John. We presented all participants with the following four rounds, in random order:

1. OVERDETERMINED POSITIVE: John drew a colored ball from each of the four urns — John won (Figure 3.5b)
2. TRIPLE-1: John drew a colored ball from urns A, B, and D, but not from urn C — John won (Figure 3.5d)
3. TRIPLE-0: John drew a white ball from urns A, B, and D, but not from urn C — John lost (Figure 3.5e)
4. OVERDETERMINED NEGATIVE: John drew a white ball from all four urns — John lost (Figure 3.5f).

Within each round, we asked participants about every singular event that featured a colored ball in the winning rounds, and every singular event that featured a white ball in the losing rounds. We also asked about every plural combination of these singulars, with the exception of four-variable plurals (we considered those questionable candidates for causal selection judgments, since they provided an exhaustive description of all drawing events in a given round) and other plurals which we considered redundant with some that we already asked. The questions were presented in random order, with no separation between singulars and plurals.

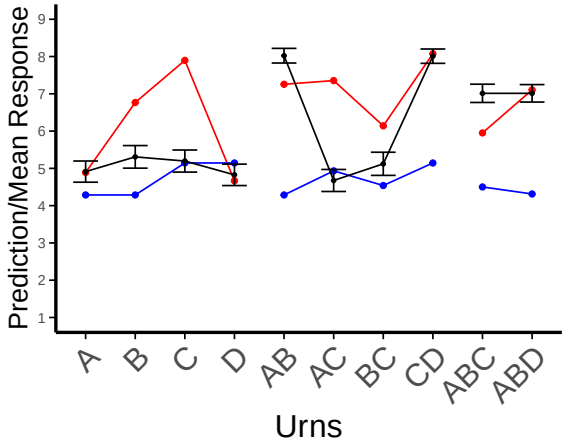
We recruited a total of 368 participants (153 male, 215 female, mean age: 37.3) from all English-speaking countries on Prolific. We excluded from analysis 57 participants who failed to correctly answer either one of our two elementary comprehension questions, yielding a final sample of 311 participants whose data we analyzed. Each participant answered all of the questions of the four conditions in this experiment.

Computational modeling

We computed the predictions of the CESM and the NSM following the same procedure as in Experiment 1. We fitted the value of s and γ for both models by finding the parameter values that resulted in the best fit between model judgments and average participant judgments across all four conditions. As in Experiment 1, we used a grid search, exploring a wide range of values for the parameter s, crossed with different values for a scaling parameter γ. For the CESM, the best-fitting value was $s = 0.21$ (with $\gamma = 0.39$). For the NSM we find $s = 0.02$ (with $\gamma = 0.28$).

We also explored a variant of the computational models that allows for the possibility that participants handle the losing cases in the non-classical fashion discussed above. We provide the details of this model in the next section.

Figure 3.6: Participants' responses, along with model predictions, for the OVERDETERMINED POSITIVE round. The red line represents the CESM predictions, the blue line the NSM predictions, and the black line represents the mean of participants' responses.



3.3.3 Results

We first go through the results for each round separately. We start each section by a brief exposition of the predictive performance of the CESM and NSM models for the round before delving into a qualitative analysis of the relevant patterns of judgments observed for that round. Note that none of the patterns we identify or the interpretation we provide for them depend on the models considered, unless explicitly specified otherwise. We provide these predictions mainly for readers interested in how state-of-the art counterfactual models fare at predicting these new data.

Winning rounds

Overdetermined positive round In this round, the player drew a colored ball from each of the four urns (as in Figure 3.5b) and therefore won the game.

Figure 3.6 summarizes our results in this condition. The CESM had a moderate but positive fit to participants' average judgments, $r(8) = .45$, while the NSM predictions were uncorrelated with participants' judgments, $r(8) = -.18$.

Participants' judgments also reveal the following patterns.

Non-linearity Participants judged that B and C were the most important singular causes. Therefore, a linear combination approach would predict that they should also view the pair $B \wedge C$ as the best plural cause. In fact, participants judged that the pairs $A \wedge B$ and $C \wedge D$ were significantly better causes than $B \wedge C$, in clear opposition to the predictions of the linear combination hypothesis.

Participants preferred non-crossing over crossing pairs There was a clear preference for pairs that did not cross the disjunction ($A \wedge B$, $C \wedge D$) over those that featured one variable on each side of the disjunction (e.g. $A \wedge C$, $B \wedge C$), (mean non-crossing: 7.02, mean crossing: 4.90; $t(df) = 23.971$, $p < 0.001$).

Weak abnormal inflation for singular variables We observed an abnormal-inflation effect at the level of singulars, meaning that participants deemed urns B and C, which contained the lowest proportion of colored balls, more important for bringing about the outcome. Formally, judgments for B and C were higher than for A and D, $t(1239.3) = -2.56, p < 0.011$. This qualitative pattern aligned with the predictions of the CESM, but not with the predictions of the NSM, which prescribed abnormal deflation in this context. No significant difference could be observed however between the two low-probability singulars, contrary to the CESM's expectations (means: 5.31, 5.19; $t(619.67) = 0.52$, $p > 0.600$).

The CESM overestimates the attractiveness of some plurals The CESM mistakenly predicted that $A \wedge C$ should be rated higher than $A \wedge B$, and $A \wedge B \wedge D$ higher that $A \wedge B \wedge C$. In both cases, the predictions come from a tendency of the model to give a very similar rating to the singular X and the pair $X \wedge Y$ if Y is a high-probability variable. This is because if $P(Y)$ is high, the correlation between $X \wedge Y$ and the outcome is very similar to the correlation between X and the outcome. This property often results in erroneous predictions, not only in this particular round, but also in the TRIPLE 1 round below, where plurals containing the variable D are overestimated. We come back to this pattern in the Discussion section for this experiment.

Triple-1 round In this round, the player drew a colored ball from urns A, B and D (as in Figure 3.5d) and therefore won the game. In such a draw, the win is not overdetermined like it was in the previous round, but clearly it is caused by the player's getting a colored ball from both *purple* urns A and B. Urn D, on the other hand, is not an active cause of the win in the present world, because it has no effect on winning in the absence of C.

Notice that, in the particular context of this round, drawing a colored ball from urn D does not simply have a low impact on the win, but in a categorical sense it is not at all a cause of the outcome in the actual world. A standard view of how causal-selection judgments work holds that only the events that can be counted as *actual causes* (Halpern 2016b) of the outcome qualify as candidates for causal selection in the first place (see for example Gerstenberg et al. 2021; Quillien and Lucas 2023). Following this logic, the causal impact score of the event "drawing a colored ball from urn D" should simply be zero, and it is unclear if plural events that contain D (such as "drawing colored balls from urns A and D") should count as actual causes or not. For simplicity, we gloss over this issue, allowing the model to give non-zero causal responsibility to D or plurals that feature D.

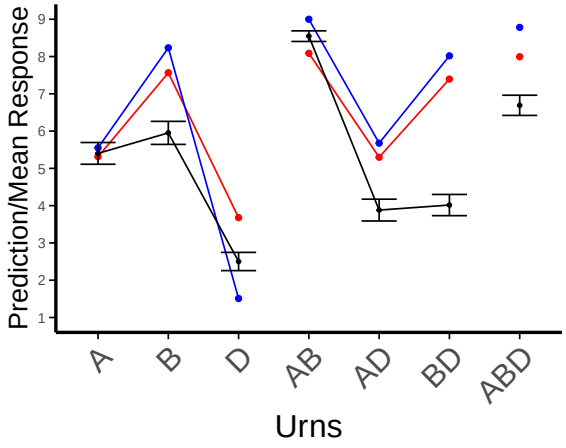
Figure 3.7 summarizes the results for the TRIPLE-1 rounds. Both counterfactual models give a good account of participants' judgments: model predictions are correlated with average human judgments $r(5) = .78$ (CESM) and $r(5) = .79$ (NSM). We now highlight the most significant patterns.

Abnormal inflation effect for singulars We did observe an abnormal inflation effect, with the low-probability urn B being ranked significantly higher than high-probability urn A ($t(617.88) = -2.5363$, $p < 0.012$), in line with the predictions of both the CESM and the NSM.

Ceiling-high ratings for the pair $A \wedge B$ Participants were almost unanimous in giving ceiling-high ratings to $A \wedge B$. Only 51 participants (out of 311) in total gave it ratings different from the maximal value of the Likert scale.

Low ratings for D, and plurals containing D Ratings for the idle variable D were very low. More than half of participants (171 out of 311) gave it maximally low ratings. Interestingly however, the ratings

Figure 3.7: Participants' responses, along with model predictions, for the TRIPLE-1 round. The red line represents the CESM predictions, the blue line the NSM predictions, and the black line represents the mean of participants' responses.



weren't as low as they were high for $A \wedge B$, suggesting that the fact that D does make a contribution to the win in other possible configurations still had some residual influence on participants' ratings.

Plurals containing D, such as the mixed pairs $A \wedge D$, $B \wedge D$, and the triple $A \wedge B \wedge D$, were systematically rated somewhere between the best cause that they contained and the low ratings of D. They were systematically rated lower than predicted by the models, which didn't penalize strongly enough the inclusion of the idle variable D. However, participants didn't seem to systematically disqualify a plural just for including the variable D (for example, by giving it ratings as low as those of D alone).

Losing rounds

The first two conditions just discussed collected judgments about the contribution of *colored ball* draws to a player's *win* in a given round of the game. The two conditions we present next instead queried participants' judgments on the contribution of *white ball* draws to a player's *loss*.

We find that, unlike in the winning rounds, participants' judgments about the losing rounds do not seem to be sensitive to the grouping suggested by the structure of the causal rule. That is, participants' judgments do not appear to be sensitive to the fact that urns A and B are on one side of the disjunction, while C and D are on the other side.

Inspired by work on plural negation in natural language, we explore one possible explanation for these patterns: when participants make causal judgments about losing rounds, they might be representing the causal rule for losing the game as LOSS $:= \neg A \wedge \neg B \wedge \neg C \wedge \neg D$ (equation 3.2), that is "you lose if you don't get any colored balls." Again, we must clarify that our idea is *not* that participants might be mistaken or confused about what the losing conditions are: we fully expect participants to understand that getting white balls from *all* urns is *not* the only way to lose the game. Rather, our hypothesis is that the under-the-hood computations of causal responsibility make use of the non-classical representation.

In order to formalize this hypothesis in a counterfactual framework, we consider a variant of our computational models featuring a parameter w, which encodes participants' propensity to represent the losing

conditions in the non-classical, language-like way depicted in equation 3.2, as opposed to the classical, normative negation of the winning conditions. Concretely, when the outcome under consideration is a loss, the subject makes a random decision in each world, where

- with probability w, the loss is determined non-classically (equation 3.2 on page 37);
- otherwise, with probability $1-w$, the loss is determined by the classical negation of the original rule, (equation 3.3 on page 37). This entails that the original models can be understood as a special case of the w models where $w = 0$;
- once it has been determined whether a given world is an instance of a win or a loss, the worlds that are not losses are recorded as wins. The impact of each variable on the models is then computed exactly as before.

We fitted the models again in this new version (using data from all four conditions), via a three-dimensional grid search (s, w; γ). The best fitting values were respectively $s = 0.21$ and $w = 0.77$ (with $\gamma = 0.41$) for the CESM, and $s = .5$ and $w = 0.77$ (with $\gamma = 1.17$) for the NSM. For simplicity, all model predictions we report use the values of the s and γ parameter fitted jointly with w, even for the base versions. Using the original fitted parameters for the base versions yields virtually identical results.

Adding the w parameter significantly improved the fit of both models, even accounting for differences in degrees of freedom (Table 3.6 on page 46). For the negative conditions below, we report both versions of the models, to showcase the impact of the new parameter.

Overdetermined negative condition The OVERDETERMINED NEGATIVE condition is the mirror image of the OVERDETERMINED POSITIVE condition. Here, the player drew a white ball from all four urns, and consequently lost, as pictured in Figure 3.5e.

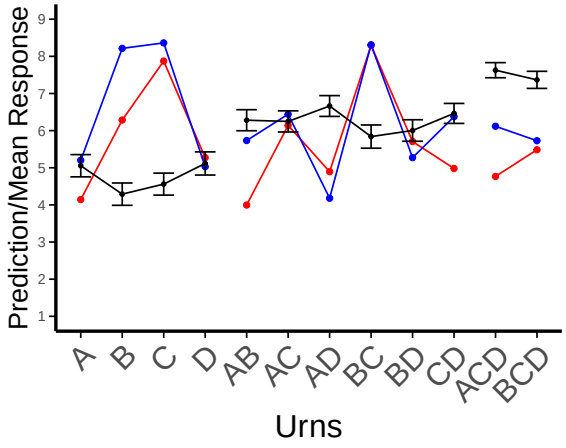
The results are summarized in Figures 3.8 and 3.9. The base versions of the CESM and NSM have a poor fit to participants' average judgments, $r(10) = -.38$ (CESM) and $r(10) = -.42$ (NSM). In contrast, the versions of the models featuring the w parameter provide a good account of the data, $r(10) = .82$ (CESM) and $r(10) = .87$ (NSM); see also Table 3.7 on page 48. We now go over our most telling findings.

Urns with the lowest number of white balls are given higher scores This effect can be observed both for singulars and for plural causes. Combinations featuring urns A or D scored higher than those featuring B or C. This pattern runs completely contrary to the predictions of counterfactual models under the classical representation of losing conditions from equation 3.3, but is captured by the version that assumes a non-classical representation of the rule.

Indeed, if participants are representing losing conditions as a disjunction of minimally sufficient conditions of that shape, we would expect their judgments to follow the logic of abnormal *deflation* and ascribe a greater causal impact to those urns out of which one is most likely to get a white ball, that is urns B and C. Instead, their judgments seem to follow a logic of abnormal *inflation*, with a preference for the urns that contain the lowest number of white balls, i.e. A, D, consistent with representing the losing conditions as a conjunction of necessary events as in equation 3.2.

No significant difference between pairs that cross the disjunction and those that do not. Participants' judgments for pairs that crossed the disjunction (e.g. A and C) were not significantly different than for pairs that did not cross the disjunction (mean crossing: 6.19; mean noncrossing: 6.37; $t(1311.9) = 1.47$, $p > 0.140$).

Figure 3.8: Participants' responses, along with model predictions, for the OVERDETERMINED NEGATIVE round. The red line represents the CESM predictions, the blue line the NSM predictions, and the black line represents the mean of participants' responses.



This finding is consistent with the idea that participants represent the rule for losing the game in the format LOSS $:= \neg A \wedge \neg B \wedge \neg C \wedge \neg D$, with no natural grouping of the variables. In contrast, a classical representation of the losing conditions would have predicted that any pair of events on the same side of the purple vs. yellow divide should be redundant, since a single white ball on either side is sufficient to cancel any contribution that this side could have made to a win. There is no such redundancy however if the representation is non-classical, where each white-ball-drawing event makes a crucial contribution to the outcome. A consequence of this pattern is an overall preference for causes mentioning more variables, with the means for singulars clustering between 4.29 and 5.12, the means for pairs between 5.90 and 6.66 and the means for triplets between 7.37 and 7.63.

Triples are rated higher than pairs Mean pairs: 6.25; Mean triples: 7.37; $t(487.6) = 8.47$, $p < 0.001$. Here again, while triples would have been redundant under a classical representation, each element of the triple makes a non-zero contribution to the outcome if the representation is non-classical.

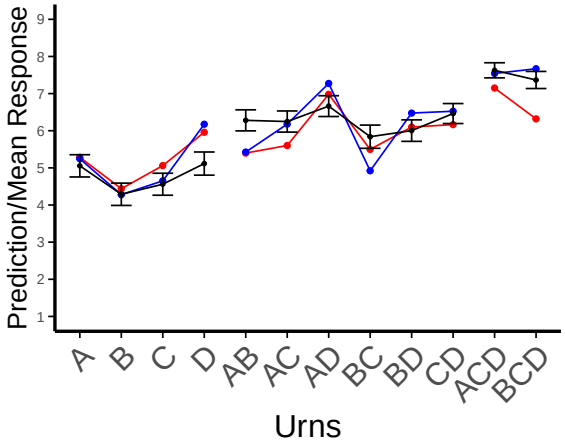
Triple 0 condition In the TRIPLE 0 round, the player drew white balls from every urn except for urn C, as in Figure 3.5f. This makes it a mirror image of the TRIPLE 1 round, where white balls are substituted for colored balls. In this round, the white ball from urn D is indispensable for the loss, whereas urns A and B are redundant with one another.

The same contrast between classical and non-classical representations of losing conditions applies in this round. Here, the w parameter that we enriched our models with encodes participants' propensity to reinterpret the rule of the game as below.

$$\text{LOSS} := \neg A \wedge \neg B \wedge \neg D$$

We take it that the non-classical representation of the losing conditions in this round is slightly different from the OVERDETERMINED NEGATIVE round because, in the actual world, a colored ball was drawn from urn C. This makes the negation of the plural entity $C \wedge D$ in our rule harder to interpret as the strong plural

Figure 3.9: Participants' responses, along with model predictions, for the OVERDETERMINED NEGATIVE round, with w parameter, encoding participants' tendency to represent the losing conditions non-classically.



negation $\neg C \wedge \neg D$, since the situation at hand is known to be one where the player in fact drew a colored ball from urn C. In other words, the player cannot possibly have lost *because* they drew a *white* ball from C, since they in fact drew a *colored* ball from C.

Results are summarized in Figures 3.10 and 3.11, and in Table 3.7 (page 48). The base versions of the CESM and NSM have a moderate fit to participants' average judgments, $r(10) = .34$ (CESM) and $r(10) = .52$ (NSM). In contrast, the versions of the models featuring the w parameter provide a good account of the data, $r(10) = .94$ (CESM) and $r(10) = .99$ (NSM); see also Table 3.7. We highlight some of the most important qualitative patterns below.

Participants prefer urns with a lower number of white balls Causal judgments for $\neg A$ were higher than $\neg B$ ($t(620) = 2.98$, $p < 0.010$), and causal judgments for $\neg A \wedge \neg D$ were higher than $\neg B \wedge \neg D$ ($t(620) = 2.34$, $p < 0.020$). The preference for urns featuring a lower number of white balls is similar to what we find in the OVERDETERMINED NEGATIVE round. Again this pattern is most coherent with a non-classical representation of the losing conditions.

The pair $\neg A \wedge \neg B$ rates higher than either of its constitutive singulars, and the triple $\neg A \wedge \neg B \wedge \neg D$ rates higher than its constitutive pairs ($t(576) = 7.77$, $p < 0.0001$). Both patterns are examples of plurals whose effect in the outcome under the classical representation is redundant with that of one of the events (singular or plural) contained within it, which should lead them to be rated at most as high as the sufficient event in question. The fact that these are rated higher by participants is again suggestive of their representing the losing conditions non-classically.

3.3.4 Overall model comparison

Table 3.6 summarizes the comparison between the models at the global level (all conditions combined). The version of the CESM that includes the w parameter has the best fit overall (BIC = 53835.36; correlation with means: $r(35) = 0.67$, $p < 0.001$). This is better than the fit of the model without the w parameter

Figure 3.10: Participants' responses, along with models predictions, for the TRIPLE 0 round. The red line represents the CESM predictions, the blue line the NSM predictions, and the black line represents the mean of participants' responses.

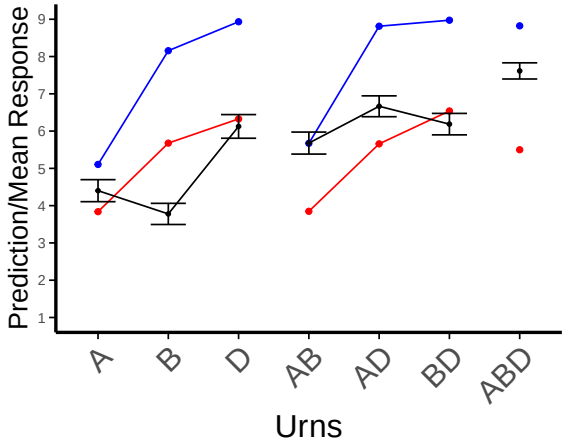


Table 3.6: Model comparisons for Experiment 2, across all conditions.

Model	LogLik	χ^2	p-value	BIC	Cor.
Baseline	-27494			54997.32	0
CESM, with w	-26905	1175.91	< 0.0001 ***	53840.57	0.67
CESM, no w	-27963	2023.44	< 0.0001 ***	55889.97	0.2609
NSM, with w	-27634	642.55	< 0.0001 ***	55295.31	0.57
NSM, no w	-28390	11511.80	< 0.0001 ***	56967.07	0.02

(BIC $= 55926.62$; correlation with means: $r(35) = 0.26$, $p < 0.001$), or than any of the versions of the NSM model (with w: BIC $= 55295.31$, cor.: $r(35) = 0.57$, $p < 0.001$; without w: BIC $= 56967.07$, cor.: $r(35) = 0.02$). In general, the versions of the models that include the w parameter are better than the versions without it, by all metrics.

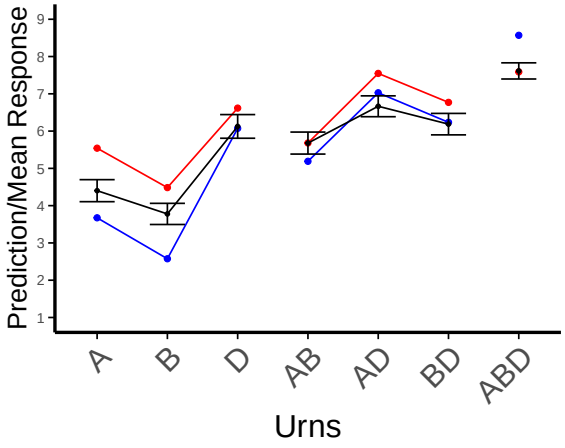
We also compared these models with a constant baseline model, which constantly made the same predictions about every question in every condition. The prediction was fitted to the data via the scaling parameter γ only. All counterfactual models had a better fit than the baseline model when assessed in terms of their correlations with mean human judgments, but only the version of the CESM that included the w parameter had a better BIC score than the baseline model.

3.3.5 Discussion

This second experiment provides more evidence in favor of the psychological reality of plural causes in the context of causal selection judgments.

Just like in the first experiment, participants' judgments for plural causes across all four rounds of the game were clearly sensitive to the probabilities attached to the corresponding events. Participants' judgments

Figure 3.11: Participants' responses, along with models predictions, for the TRIPLE 0 round, including the w parameter. The red line represents the CESM predictions, the blue line the NSM predictions, and the black line represents the mean of participants' responses.



in the OVERDETERMINED POSITIVE round corroborate the non-linearity between participants' judgments for plurals and their judgments for the singular causes that constitute them. Given the pattern of abnormal inflation observed for singular variables, favoring B and C over A and D, a *linear* reconstruction of participants' judgments for plurals would have us expect the pair $B \wedge C$ to rank above all others pairs, when in fact it ranks much lower than the $A \wedge B$ and $C \wedge D$ pairs.

The winning rounds of the experiment also demonstrate that plural causes featuring more variables are *not necessarily* rated higher than proper subsets of the variables they contain. The TRIPLE 1 round shows a clear pattern in this regard: every time a plural features the variable D, its rating is systematically lower than that of the same cause (singular or plural), minus the variable D. This contradicts the hypothesis that adding more variables always makes an explanation more attractive, which the results from Experiment 1 could not rule out. And the phenomenon is not limited to the situation where an idle variable like D features in a plural: a similar observation can be made about the triplets $A \wedge B \wedge C$ and $A \wedge B \wedge D$ in the OVERDETERMINED POSITIVE condition, both of which are rated lower than the best pair that they contain, $A \wedge D$. Thus, although plurals featuring more variables might be descriptively more thorough, they can still be unappealing if their overall counterfactual dependence profile drops as a result of the variables added.

We also uncover properties of plural causal judgments that go beyond what is expected based purely on patterns of counterfactual dependence. First, in the winning rounds of our second experiment, participants dislike causal explanations that "cross" the disjunction $(A \wedge B) \vee (C \wedge D)$, above and beyond what is predicted by counterfactual models. The fact that the causal rule features two clearly distinct sufficient conditions seems to exert an influence on participants' explanatory preferences not fully captured by the counterfactual dependence profile of the variables in question.

Second, the following property of the CESM was not reflected in participants' judgments. The model tends to give a very similar rating to a singular X and the pair $X \wedge Y$ if Y is a high-probability variable. This is because if $P(Y)$ is high, the correlation between $X \wedge Y$ and the outcome is very similar to the correlation between X and the outcome. This property often results in erroneous predictions, like in the OVERDETERMINED POSITIVE round where the model predicts (against participants' judgments) that the pair

Table 3.7: Table of model fits per model and condition. The *Cor.* column indicates the item-level correlation between model predictions and mean participant responses per question.

Condition	Model	BIC	AIC	Cor.
OVERDETERMINED POSITIVE	CESM	15009.14	14991.01	0.45
	NSM	15278.45	15260.32	-0.18
TRIPLE 1	CESM	10335.21	10318.16	0.78
	NSM	10505.71	10488.66	0.79
OVERDETERMINED NEGATIVE	CESM, no w	19238.14	19225.69	-0.38
	CESM, w	17731.07	17718.62	0.82
	NSM, no w	19164.41	19151.96	-0.42
	NSM, w	17703.51	17684.84	0.87
TRIPLE 0	CESM, no w	10853.57	10842.19	0.34
	CESM, w	10335.21	10318.16	0.94
	NSM, no w	10655.7	10644.33	0.52
	NSM, w	10422.88	10405.82	0.99

$A \wedge C$ should rate higher than the pair $B \wedge C$, and the triplet $A \wedge B \wedge D$ higher than $A \wedge B \wedge C$.

There is likely more than one way to resolve this discrepancy. One avenue we think is worth exploring is to re-examine the assumptions we have made about the way people simulate alternatives to a plural event when they judge whether that plural event caused E. Here we made the conservative assumption that people sample a counterfactual alternative to the plural event by using the same procedure they use to sample the *background variables* in the causal system (i.e. the variables that are not the current focus of causal judgment). Under this assumption when people judge whether a plural like $A \wedge C$ caused E, they sample alternatives to $A \wedge C$ in a way that is sensitive to the probabilities of both A and C. If $P(A)$ is high, then re-sampling $A \wedge C$ tends to have very similar effects as just re-sampling C, in that most of the $[\![A \wedge C]\!] = 0$ worlds will be $[\![C]\!] = 0$ worlds. Future research should explore the possibility that people in fact re-sample the candidate plural cause in a different way.

Finally, we found that counterfactual models could only account for participants' judgments in the losing rounds if we assume that participants handle the losing conditions, in their internal computations, in a way inconsistent with the classical-logical negation of the winning conditions. Specifically, participants seem to be representing the negation of the winning conditions (i.e. the losing conditions) in a way consonant with how natural-language represents the negations of plurals. This hypothesis is supported by the observed preference for urns with a lower number of white balls and for plurals containing a higher number of variables, and by the absence of a preference for plurals that cross the $(A \wedge B) \vee (C \wedge D)$ disjunction in the negative rounds of the experiment.

3.4 General discussion

Humans make systematic judgments regarding which of several events influencing an outcome should be considered as *the cause*, or the most important cause of that outcome. These *causal selection* judgments are the object of a rich and actively expanding section of the psychological literature on actual causation. So far,

however, that literature has been exclusively focused on *singular* events, identified with the distinct nodes of the relevant causal system. In this article, we argue that its scope should be extended to include *plural* events, featuring multiple variables.

Our experiments present strong evidence that judgments about plural events cannot be captured in terms of linear combinations of the judgments for the events that constitute them. There appears to be no obvious way of combining participants' causal judgments regarding any two events A and B that would predict their judgment for the event "*A and B*." Our results thus establish the psychological reality of plural causes: plural causes are treated by the mind as causal entities in their own right, and their impact on the outcome is apprehended in a *holistic* fashion. We also uncovered patterns in participants' judgments that are difficult to explain under a naive view of what a plural is and how it interacts with negation, but are readily accounted for under the *sui generis* yet mathematically rigorous theories of plurality from natural-language formal semantics.

3.4.1 Summary of our findings and their immediate consequences

Plural cause judgments cannot be reconstructed as linear combinations of singular judgments.

It seems *prima facie* plausible that, when people make a causal judgment about whether "A and B caused E," they might judge how much A caused E, judge how much B caused E, and then combine these two judgments into a single judgment for the plurality. Under this view, plural causal selection would be entirely predictable from facts about singular causal selection. One of our main goals was to rule out this null hypothesis.

In our two experiments we designed situations in which computational models predict that, if plurals are processed in a holistic manner, judgments about plural causes should not be simple combinations of judgments about their constituent singular variables. Participants' judgments in these situations supported this prediction.

This finding is key to establishing the relevance of our object of study. Were participants' evaluation of "*A and B*" systematically proportional to their evaluations of A and B taken separately, there would be no reason to study judgments specifically about plurals, or to build theories around such judgments.

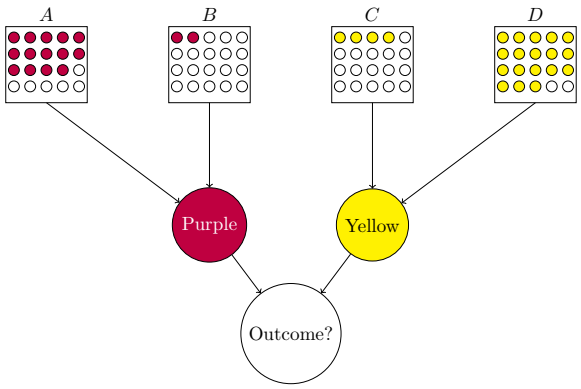
Counterfactual models can account for plural causation judgments

A growing body of research provides strong evidence that causal judgment involves counterfactual thinking (e.g. Gerstenberg et al. 2017; Kahneman and Miller 1986; Krasich, O'Neill, and De Brigard 2024; Quillien and Lucas 2023). At the same time, there are debates about what phenomena counterfactual theories can explain (Hall 2004; Henne 2023; Lombrozo 2010; Rose, Sievers, and Nichols 2021; Sytsma 2020), and about the computations that counterfactuals might be an input to (Icard, Kominsky, and Knobe 2017; Quillien 2020).

Our experiments provide a rich opportunity to probe the out-of-distribution generalizability of counterfactual theories. None of the counterfactual theories that we are aware of were developed with the goal of explaining data about how people make plural causation judgments. Consequently, accounting for these judgments off-the-shelf would constitute important evidence in favor of these theories.

We found that two recent counterfactual models of causal judgment (Icard, Kominsky, and Knobe 2017 and in particular Quillien and Lucas 2023) can quantitatively account for many features of participants' judgments. In particular, when participants' judgments about plurals diverge from a linear combination of their constituent singular causes, they typically do so in the way that is predicted by the counterfactual models. As such, our results strengthen the case for counterfactual theories.

Figure 3.12: Causal graph representing participants putative model of the situation.



Yet, both counterfactual theories failed to predict the shape of people's judgments in the rounds of the game where the player *loses*. Counterfactual theories are only able to account for these data if we make an additional assumption about how participants might represent the causal structure when evaluating counterfactuals: they use representations with properties related to those found in the representations of natural-language plurals. The plausibility of a counterfactual account of these data thus depends on the plausibility of this additional assumption, which we address in detail later in this section.

Causal judgments favor sets of variables belonging to the same disjunct

In the game that participants played in Experiment 2, the player needed to draw either two purple or two yellow balls in order to win the game. From a logical point of view, this rule is a *disjunction*: the player wins if either one of the conditions for victory is met; each condition is a *disjunct*.

Participants favored plural causes that do not cross the boundary between the two disjuncts. Suppose for example that the player drew purple balls from urns A and B and yellow balls from urns C and D. In this situation, participants would be reluctant to say that the player "won because he drew a colored ball from urn B and urn C." The counterfactual models also disfavored these boundary-crossing plurals, but participants did so to an even greater extent than predicted.

There are different possible explanations for this pattern. At a superficial level, for example, participants might have preferred causal explanations that mentioned balls of the same color because of low-level perceptual biases.

At a deeper level, participants might have built an internal representation of the game in terms of a causal model with a particular structure. This causal model would contain intermediate variables (in the technical, causal-model sense of "variable," Pearl 2000) representing whether each condition for victory (getting two purple balls) and (getting two yellow balls) is met, see Figure 3.12. Such a model would be distinct from one without the intermediate variables, in that it would support new kinds of interventions and therefore inferences unavailable to the more straightforward model. It remains to be seen whether this difference in intervention potential makes the right predictions.

Another possibility is raised by research on the mental representation of disjunction (Chung et al. 2022; Koralus and Mascarenhas 2013; Walsh and Johnson-Laird 2004). Studies of deductive reasoning have investigated how people reason about logical statements of the shape $(A \wedge B) \vee C$. In experiments replicated

and varied multiple times, participants overwhelmingly conclude B from the two premises $(A \wedge B) \vee C$, and A (Koralus and Mascarenhas 2018; Picat and Mascarenhas 2024; Sablé-Meyer and Mascarenhas 2021; Walsh and Johnson-Laird 2004). But this is a fallacy: it is compatible with the premises but not the conclusion that A and C should be true while B is false. Koralus and Mascarenhas (2013) explain this fact in terms of question-answer dynamics: the disjunction in the first premise is naturally interpreted as demanding the participant *choose* between one of the two disjuncts. This in turn induces dependencies between propositions occurring *within disjuncts*: in the context of $(A \wedge B) \vee C$, the second premise A is seen as an answer in the $A \wedge B$ direction, introducing dependence between A and B. In general, this approach predicts that the conjuncts within each disjunct will be taken, as it were, to *hang together* in a cohesive way, so that learning about one will constitute evidence in favor of all of the others.

There is even evidence of such effects absent the language of disjunction, in experiments where the same information was conveyed by means of visual stimuli in the form of animations (Chung et al. 2022), indicating that this "packaged" way of representing a disjunction is not simply a fact about the interpretation of the word "or" and its equivalent locutions. Rather, these rich, structured disjunctive representations which induce dependencies not predicted by standard Boolean interpretations of logical connectives are available to human minds far more generally. In particular, they might have been available to participants in our Experiment 2, and may have played a part in shaping their causal judgments, by pushing them to associate A and B on the one hand and C and D on the other more tightly than is predicted by classical accounts of disjunction, whether deductive or probabilistic.

Plural causes featuring more events are not necessarily better

In Experiment 1, we found that participants preferred causal explanations that mentioned the most causes, and that this preference was stronger than predicted by counterfactual models. We probed the extent of this trend in Experiment 2, where we found that mentioning more causes does not always make a causal explanation better. For example, a causal explanation mentioning only two events A and B might be judged better than an explanation mentioning A, B, and C.

These results suggest that causal judgments are subject to a trade-off between two different considerations. On the one hand, people might favor explanations that give detailed information about what events happened. Since every explanation of the shape "*X happened because Y*" comes with the implication that "*Y happened*" (Halpern 2016b), causal explanations that feature many causes offer more complete descriptions of what happened. With respect to this criterion, plural cause explanations are always more helpful than singular ones, since they highlight more true facts about the situation.

On the other hand, causal explanations convey information about patterns of counterfactual dependence (Quillien 2020). Under this criterion, large plural causes can sometimes be worse. For example, an explanation mentioning three events A, B, and C might misleadingly suggest that the outcome strongly covaries with the conjunction of these three events, across counterfactuals.

In losing rounds, participants appear to simulate counterfactuals using a different representation of the rule

In the Experiment 2 trials where the player loses the game, we found it is difficult to account for the data if we assume that participants internally represent the conditions for losing the game as the classical complement of the conditions for winning. Instead, judgments can be captured quite adequately if we suppose that they simulate counterfactuals using a different representation of the rule of the game, in the particular case of

losing rounds. Specifically, participants seem to represent the losing conditions as "the player loses the game if they draw a white ball from all urns," a representation not justified by standard Boolean logic.

The representation our participants seem to have for the losing conditions is however quite consistent with fundamental facts about the semantics of plurals. In natural language, the negation of a plural event is naturally represented as the negation of each of its singular constituents, rather than the negation of their conjunction. That is, for a sentence like "Mary and John didn't come to the party," we infer that Mary did not come and John did not come, rather than that at least one of them did not come (Krifka 1996; Lappin 1989; Löbner 2000; Szabolcsi and Haddican 2004). In our experimental setting, negating the conditions for winning the game yields the following losing conditions expressed in natural language: "you lose if you don't get the purple balls and you don't get the yellow balls." Natural-language plural negation then predicts that these losing conditions correspond to not getting *any* purple balls and not getting *any* yellow balls. Consequently, our results provide suggestive evidence that a signature effect of the semantics of plurals, often dubbed *homogeneity* in the linguistics literature, affects people's representations of events.

It is important to be more precise about the exact level at which we take this mental representation of the rule to occur. We are not claiming that our participants explicitly believed that one needs to draw a white ball from *every single urn* on the screen to lose. We did not directly probe participants' judgments on their understanding of the rule, but they played the game of chance for 10 rounds before responding to any causal-judgment questions, and these draws included losing cases in which colored balls were drawn. Some additional controls could be run (such as making the task simpler by equalizing the probabilities of the urns) to ensure that the effect does not stem from people's confusion about the parameters of the task. Moreover, the effects we find are unlikely to stem from linguistic experimenter demands when interpreting the causal statements verbally. We asked participants about the causal impact of "drawing a white ball" on a "loss," never about the impact of "not drawing a colored ball" on "not winning."

Instead, we propose that a non-classical representation of the losing conditions is deployed when participants implicitly simulate counterfactual possibilities. In other words, the effects we find stem from features of a (likely unconscious) process of counterfactual simulation, rather than as explicit and conscious misconceptions about the losing conditions of our game of chance.

3.4.2 Broader theoretical implications: natural-language semantics and causal cognition

Two of our proposals were inspired by work on natural language, specifically work on the formal semantics of plurals. This work inspired our proposals that i) plural causes are processed holistically, and ii) people deploy the equivalent of natural-language plural negation when simulating counterfactuals for losing events.

Why should work on natural language semantics be relevant to causal judgment? Here we present several considerations which we mean neither exclusively nor exhaustively.

First, participants in our experiments have to use natural language to read the description of the causal structure (i.e. the rules of the game) and the causal statements they have to evaluate. This stage of linguistic processing might "package" information into a particular representational format. For example, when participants read "the player won the game because he got a colored ball from urns A and B," they might process the event "he got a colored ball from urns A and B" as one holistic entity, as is typically the case according to linguistic theories of plurality (Link 1983). The causal judgment process then treats the plural event as a holistic entity because it received the input in that format. Consequently, it's not that participants' causal reasoning is manipulating language-like representations, it's simply that the language suggested some degree of togetherness between the variables mentioned, in virtue of its plural semantics, but the

causal-reasoning system now handles this togetherness in its own proprietary way.

In favor of this hypothesis is the fact that our basic experimental findings regarding holistic evaluations of plural causes are largely predicted by the two causal-selection theories we considered. Both causal-selection theories, as instantiated in our models, had a good fit with participants' judgments, especially in Experiment 1. On the other hand, our Experiment 2 provides evidence that subjects' judgments traffic in plural representations even when those are not directly prompted by the instructions of the experiment, in ways that go beyond the predictions of the causal-selection theories at hand. In our losing rounds, participants had to internally compute the conditions for *not winning*, in order to assess the extent to which "the player lost because" of this or that cause. Now, the conditions for *winning* were given linguistically and included two plurals connected by a disjunction ("colored balls from urns *A* and *B* or colored balls from urns *C* and *D*"), but crucially our instructions never presented the *negation* of this disjunction of plurals. Accordingly, and unlike the fundamental facts about holistic entities, the causal-selection theories in and of themselves accounted for the losing rounds very poorly indeed.

This more surprising fact can be interpreted in at least two different ways. One possibility is that some participants might be running an internal monologue when completing the task. That is, subjects might be *talking to themselves* in the course of their attempts to put together a representation of the relevant causal structure. For example, in the conditions where the player loses the game, they might be reconstructing the conditions for losing by saying to themselves "you lose if you don't get the purple balls and you don't get the yellow balls, this means that you lose if you don't get any colored balls."

Such a proposal is in principle testable; for example, verbal shadowing tasks can interfere with reasoning processes that plausibly rely on internal natural language use (Carruthers 2002). If such self-talk is the cause of the signature effects of plurals that we identified here, such as subjects' non-classical reconstructions of the losing conditions, we would expect these manipulations to bring subjects' behavior back in line with the classical representation of the negation of the winning conditions. While we think such a study is worth doing, we are skeptical of the prediction, which is that participants would have been perfectly capable of handling this complex rule had they not been engaging in deliberate reasoning.

Another hypothesis, which we favor, is that our judgments about causes display the same sort of effects found in language because the underlying cognitive processes operate on the same kinds of representations as our language faculty itself. That is, both natural language and causal judgment might depend on a shared *language and logic of thought* which supports many of our higher cognitive faculties.

These considerations resonate with the recent renaissance which Jerry Fodor's (1975; 2008) *language of thought* hypothesis has been enjoying. As Quilty-Dunn, Porot, and Mandelbaum (2023) recently observed, much current research embraces the idea that human cognition relies on symbolic representations of a language-like nature, of the kind that Fodor proposed were at the core of human thought.

While the classical illustrations of language of thought came chiefly from the domain of natural language and general purpose, integrative thought, current research on this program has been paying particular attention to areas of cognition that are minimally connected to language, if at all, and plausibly do not require integrative thought. Part of the reason for this is sound methodology: as noted by Quilty-Dunn, Porot, and Mandelbaum (2023), those investigations provide a new class of arguments in favor of language of thought as a general hypothesis, showing how language-like representations might pervade cognition across the board, and in plausibly domain-specific ways.

The work presented here has a mixed status in this regard. On the one hand, our participants seem to make use of language-like representations in a cognitive process that does not inherently depend on natural language, namely counterfactual sampling and causal judgment. Indeed, no extant theory of causal judgment even suggests that the phenomenon might hinge on language in any appreciable way. This might suggest that

we are seeing here yet another proprietary language of thought: a language of thought for causation.

On the other hand, causal reasoning applies to every walk of human life, from ecologically natural contexts like tool making or animal husbandry, to highly abstract modern contexts such as science or public policy. Given that we are arguing for representations entirely parallel to linguistic representations, the unbounded general-purpose representational system *par excellence* (Chomsky 1965; Hockett 1960), we are inclined to think that causal reasoning, whether computed by the general-purpose reasoning system or by a causation-specific system, taps into the language of thought in the broadest sense: general-purpose, integrative thought.

Our perspective also recovers Fodor's view that, significant differences between natural language and thought notwithstanding, natural language itself is an important tool for investigating thought (Fodor 2001), in the sense we are concerned with here. We embrace this approach and would add that specifically *formal natural-language semantics* constitutes a particularly fruitful source of hypotheses about the most general-purpose language of thought. The present study can be taken (among other things) as a case study illustrating the power of this methodology. The formal semantics of plurals allowed us to consider novel hypotheses about causal cognition. The result is a contribution not just to our understanding of causal judgment, but potentially also the representational structure of higher cognition more generally: alongside the nigh-universally recognized standard Boolean operations, human thought might involve the more general and more expressive lattice-theoretic operations found in linguists' theories of plurality. Crucially, this contribution gets to co-opt the mathematical rigor that characterizes the linguistic work which inspired it, providing not just a *language* of thought, but also a *logic* and a *model theory* of thought.

In conclusion, we think that the time is ripe to formulate strong and mathematically rigorous hypotheses about the representational arsenal of human thought. Formal natural-language semantics offers a largely untapped fountain of such hypotheses, with few though notable exceptions (see in particular Phillips and Kratzer 2024, for one of our favorites). We are sympathetic to the view that there may be many domain-specific languages of thought (Mandelbaum et al. 2022; Sablé-Meyer et al. 2022), but we find that currently there is far more that can be done and in fact has already been done on the domain-general language of thought. Unlike its special-purpose relatives, the general language of thought has been the exclusive focus of investigation for the past fifty years on the part of a small but dedicated community of linguists, who have been fastidiously building an impressively broad and deep body of mathematically precise scholarship on the model theory and the logic of general language of thought in this sense. Many if not most would be loath to characterize their work this way, and would likely insist that their theories apply, at best, to *language*, thought being the purview of the philosophers and the psychologists. We disagree, and we believe that the present work illustrates how, if the language of thought is at least one of the best games in town to cash out a computational and representational theory of mind, then the *formal* study of natural language *meaning* offers the most productive and most woefully underexplored path toward building a mathematically rigorous theory of domain-general mental representations.

Chapter 4

A neural approach to causal selection judgments

4.1 Introduction

In the previous chapter, we explored people's causal selection judgments for multivariate causes. We highlighted, among other things, two key patterns of judgments. First, we observed that the disjunctive structure of the rule presented to subjects had an impact on their judgments. When presented with a rule of the form

$$E := (A \wedge B) \vee (C \wedge D) \tag{4.1}$$

people judged that conjunctions of variables like "A and B", that sit on the same side of the disjunction, were better explanations for why a player won a round of the game than conjunctions that crossed it, like "B and C" or even "A and B and C". The preference for the former conjunctions went above and beyond what was explained by their counterfactual dependence relation to the outcome alone.

Second, when people were tasked with explaining why a round of the game was *lost*, they showed no particular preference for the explanations "A and B" and "C and D", nor did they exhibit an opposite preference for pairs featuring one white ball from each side of the disjunction —– as one could have expected them to do, were they tracking the minimal sufficient conditions to lose:

$$\neg E = (\neg A \wedge \neg C) \vee (\neg A \wedge \neg D) \vee (\neg B \wedge \neg C) \vee (\neg B \wedge \neg D) \tag{4.2}$$

Instead, they favored explanations mentioning more variables generally, including when some of these variables were redundant with one another for satisfying the losing conditions. They also did not show a preference for the urns containing more colored balls, as every theory based on the assumption that they were tracking the losing conditions in 4.2 expected them to do.

These facts are surprising under counterfactual theories based on structural causal models. This is because they operate on a possible world semantics (Halpern and Pearl 2009; Pearl 2000), meaning that the counterfactual relations captured by SCM equations are to be understood as relations between worlds. For example, a counterfactual statement like:

$$P^{\mathscr{M}}(E = 1 \mid C = 0, \text{do}(C = 1)) = 1$$

says that in all of the worlds compatible with a model $\mathscr{M}$ where C didn't occur, but where we intervene to make it occur, E occurs.

A natural consequence of this semantic postulate is that equations in an SCM should obey the logical principle of Substitution: replacing any proposition in a SCM equation by another alternative proposition that tracks the same set of worlds should preserve all relations entailed by that model. This means that the rule in (4.1) could have equivalently been written in the conjunctive form

$$E := (A \vee C) \wedge (A \vee D) \wedge (B \vee C) \wedge (B \vee D)$$

From this perspective, it is puzzling that people should care about disjunctive form of a rule. For the disjunctive nature of the rule in (4.1) is but a notational choice we make as we model subjects' causal knowledge. Hence it cannot be part of our account of how people generate causal explanations off of that knowledge.

In this chapter, I propose that the internal simulation models we use to generate counterfactuals and formulate causal explanations possess more structure than SCMs and their classical semantics typically capture. To make this proposal clear, I show in this introduction how it can be articulated from two different but complementary perspectives: in terms of the representations or mental tokenings that we use as we compute explanations for some fact, or in terms of the mental devices or programs that underly those computations. From one point of view, the program perspective is more fundamental. Everything in this chapter can in some way be understood as spelling out consequences of the fact that people generate counterfactuals and compute explanations using some concrete internal generative device rather than read them off a descriptive model (Icard 2017). As I will try to make clear below, what the perspective that cares about internal programs adds to our understanding of causal explanations is an acknowledgement of the extra structure that these programs possess and that it is typically glossed over by descriptions of our causal knowledge in the formalism of structural causal models. At the same time, rigorous hypotheses about that extra structure can only be appropriately formulated by going through the level of analysis provided by theories of mental representation.

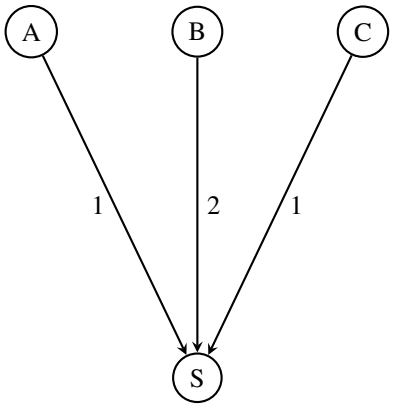
Structural models vs. programs

Let's start by illustrating the difference between the structural model that captures a certain functional relation between variables and the concrete device by which one computes the same function via a simple example. Consider the following weighted sum equation:

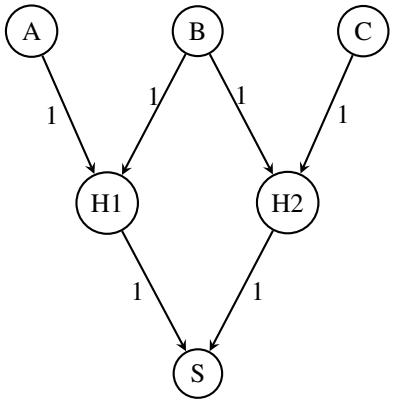
$$S := A + 2B + C \tag{4.3}$$

The relation between A, B, C and S described by (4.3) can be computed by various devices. The networks represented in Figure 4.1 instantiate two possibilities, in the form of two neural models. In both networks, each node computes a sum of the inputs from its parent nodes, weighted by the value on the edge that connects them to it. In the first one, the output node is directly fed the inputs A, B, C, whereas, in the second one, the contribution from B is broken down into two components: one in which it is put together with the contribution from A, the other in which it is put together with the contribution from C, before they are summed together into S.

It is easy to see that each of these representations is going to capture the exact same relations between variables A, B, C and S, even if we allow for interventional queries (which are trivial in this case since A, B, C have no parents in the graph). And yet there is a clear sense in which they are structurally different. The difference is not about the functional relation between A, B, C and S that they encode, but in that they force the data flow going from inputs to outputs through different routes. The availability of several options for



(a) Network N_1: A, B, C send inputs directly to S.



(b) Network N_2: Two hidden nodes mediate between inputs and S.

Figure 4.1: Each neuron computes a weighted sum of the inputs it receives from its parents, where the weight on each edge multiplies the corresponding input.

encoding the same relationship is also not specific to the use of neural representations. It would be easy to construct analogs of the networks in 4.1 in terms of cells of an Excel-like spreadsheet, for example, or even in terms of mechanical abacuses. In each case, two different options for representing the same thing may present different inner structures.

SCMs obviate that same structure because they are about describing the relations that hold between observable events, not the mental procedures by which those relations are computed; because of that, they would typically contain only the minimal amount of structure required to answer every observational and interventional query about the domain — a principle sometimes made explicit under the heading of "minimality" in Pearl (2000) and Pearl and Verma (1991). If equation (4.3) was made part of a structural model (with variables A, B, C tied to some exogenous distribution), it would induce a graph with only as many nodes as N_1.

My contention is that people's representation of causal relationships in contexts involving boolean rules like the ones studied in the experiments presented in the previous chapter (and in a large part of the literature on causal judgment) are also non-minimal in that sense, and involve intermediate layers between inputs and outputs in an analogous way. On top of that, I argue that these intermediate layers play a non-trivial role in the processes by which we generate causal explanations. They play a crucial role both in the process by which we simulate counterfactuals and in how we assign causal importance to events as a function of the counterfactuals we simulated.

Before I look at that role in detail, a question to address is: what reasons do we have to assume that our inner programs for representing boolean functions are non-minimal. Here again an analogy with the summation devices presented in Figure 4.1 will help. One advantage that the strategy instantiated by N_2 has over that of N_1 is that the former does not require the use of a multiplication operation. It decomposes the product in N_1 into additive operations at each gate (with all weights having value one, the product can be simplified away). So that although it passes signals through more convoluted routes, N_2 can operate with a more restricted class of functions.

The situation is essentially similar when it comes to the representations of boolean functions. At a

neural level, encoding arbitrary Boolean functions requires at least one intermediate (hidden) layer between inputs and outputs once we impose the restriction that the network's computations are based on parametric, differentiable functions with continuous weight parameters. In such networks, we rely on linear threshold units (or continuous relaxations thereof) to implement discrete decisions. However, a well known observation in the theory of neural networks (Minsky and Papert 1969) is that to capture all Boolean functions with such units we need at least one intermediate layer between inputs and outputs. There are many reasons why imposing this parametric constraint makes sense. The most obvious one is the higher physical plausibility of such functions. Another one is the fact that differentiable functions offer a better basis for error-driven learning procedures by means of which we can acquire causal knowledge from examples. Chapter 7 of this dissertation cashes out the explanatory purchase afforded by the use of such functions for a theory of causal learning from explanations.

Now, this constraint on functional form does not impose requirements beyond having at least one intermediate layer. The theory that I will propose in this chapter assumes more precise constraints about the architecture. Specifically, it assumes that people represent Boolean functions like that in (4.1) via a three-layer neural model N like that represented in Figure 4.2 and containing three layers as follows. The **input** layer of N contains nodes representing the arguments A, B, C, D that serve as inputs to the function. The **output** layer of N contains a single node representing the output E, whose value is determined by the function. The intermediate **hidden** layer contains exactly one node for each minimal sufficient condition represented in the proposition's disjunctive normal form (DNF). Concretely, to capture $(A \wedge B) \vee (C \wedge D)$, we include one hidden neuron for the conjunction $A \wedge B$ and another for the conjunction $C \wedge D$.

Each hidden neuron thus "tracks" one of these conjunctive clauses: it is connected only to the relevant input nodes (e.g. the $A \wedge B$ neuron receives inputs from A and from B) and the weights on these connections and the threshold on the hidden neuron are such that it activates if and only if both are active. Each hidden neuron is also connected to the output neuron; this time the weights and thresholds are such that the output neurons activates if any of them is active. We look more precisely into the relevant weights and activations functions in the sections to come. For now, we can take a bird's eye view of them and just observe that the input-to-hidden connections implement conjunctive gates through continuous weights and activation functions, while the hidden-to-output connections implement a disjunctive gate. In that sense, this choice of architecture emphasizes the minimal sufficient conditions for making E true. Its hidden layer structure mirrors the disjunctive form of the rule it implements.

This disjunction-centric architecture is not in itself justified by the fact that computations are carried over a neural model. It is possible to represent boolean functions in three-layer networks in other ways — to say nothing of networks with more layers. Another salient systematic strategy, for example, would be to have the hidden layer mirror the conjunctive normal form instead. The focus on sufficient conditions finds its justification elsewhere, in symbolic theories of mental representations. It echoes an emphasis on *alternatives* in mental model theories of reasoning (Johnson-Laird 1983a) and inquisitive semantics (Groenendijk 2008; Mascarenhas 2009). For example, Koralus and Mascarenhas (2013) take the representation of the relation in (4.1) to involve the representation of two alternatives $\{a \wedge c,\ c \wedge d\}$. Each of these alternatives tracks some minimal set of facts about the world that suffice to make E true. This way of representing relations is a consequence of the impossibility for limited beings to track all of the worlds compatible with a proposition — as we would have to assume they do if we were to take possible-worlds semantics literally, as a theory of mental representations. Those limits impose that we grasp the meaning of a proposition not by tracking all of the contexts in which it holds true, but only the minimal sets of conditions sufficient to ensure its truth.

In **Section 4.2** of this chapter, I develop these ideas by introducing a semantic framework for declarative logic programs (Lloyd 1984; Robinson 1965). Logic programs are a class of declarative programs that consist

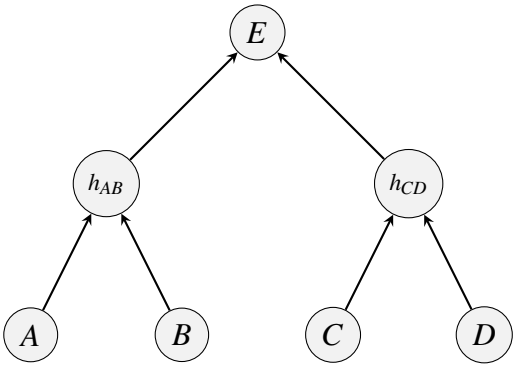


Figure 4.2: A three-layer neural network encoding $E = (A \wedge B) \vee (C \wedge D)$. Input nodes represent propositional variables (A, B, C, D), hidden nodes represent minimal sufficient conjunctions ($A \wedge B$ and $C \wedge D$), and the output node computes the final disjunction. Each hidden neuron performs a thresholded AND; the output neuron performs an OR over those hidden neurons.

in sets of Horn Clauses of the form $(\text{Head} \longleftarrow \text{Body})$, such as:

$$E \leftarrow A, B \tag{4.4}$$

$$E \leftarrow C, D \tag{4.5}$$

where each clause is to be read as a goal-directed instruction: "To prove **Head**, satisfy all of the conditions in **Body**". I explain how under some assumptions regarding the proof procedures by which we grasp the relations between propositions, these programs make sense of the emphasis on sufficient conditions in terms of computational constraints. The body of each clause in a logic program is a list of propositional atoms or literals whose conjunction constitutes a sufficient condition for proving the head. Clause (4.4) for example, allows us to prove E by first proving A and B (via other clauses in the same program). This simplifies computation in that once we have done so, we no longer need to explore other routes to E such as the one mapped out by (4.5). Logic programs modularize representations at a different level than SCMs, which is not that of individual variables but of provability pathways. My argument is that this way of modularizing representations recoups the independent observations by mental model theories summarized above. I engage with logic programs as a means to capture facts evidenced by theories of mental representations in a different framework, that puts an emphasis on executability conditions.

One can see at a very high-level, how the modular division of labor entailed by the clauses in (4.4)-(4.5) might explain subject's preference for explanations that sit on the same side of the disjunction $(A \wedge B) \vee (C \wedge D)$. "$A$ and B" is a very straightforward explanation for E. It does exactly as much work as is needed to prove E via clause (4.4), and it carries all of the procedure through that same clause. By contrast, "B and C" or "A, B and C" recruit two different clauses on their way to the outcome, to more or less the same effect.

The logic programs perspective also sheds light on why the same effects of logical form are *not* observed when people are tasked with explaining negative outcomes. In standard logic programs (more specifically, in the class of *general logic programs* to be introduced in section 4.2), proving the negation of a propositional atom E is not done directly, say by resolving a clause with $\neg E$ as its head. Instead, the negation of E is understood as *failure* to derive E (noted $\sim E$) from the clauses of the program. Failure to derive is more costly than positive derivation. Positive derivation amounts to an existential proof: to show that there is a clause in

the program from which E can be derived. Once one such clause is found the procedure can terminate. To fail to derive, on the other hand, amounts to a universal proof: it involves showing that there *isn't* a clause in the program from which E can be derived. As a result it involves trying out *all* of the clauses in the program that have E as head, and make sure each of their satisfaction condition cannot be met. Again, one can see how this recoups observations as to how we may check the falsity of a proposition on the basis of models that track its verification conditions. To show that E does not hold based on a representation of the alternatives $\{a \wedge c,\ c \wedge d\}$ that would make it true, I have to check that *both* sets of satisfaction conditions are not.

This — again at a high-level for now — explains why, when tasked with explaining a negative outcome, people didn't penalize explanations that involved more variables than necessary to explain the outcome: if any attempt to explain a loss involves checking *all* of the ways in which one could have won anyway, then we expect the processing costs to be equalized across the different possible explanations. This in turn favors more prolix explanations, because mentioning more relevant fact in one's explanation of the outcome comes at no extra cost.

Logic programs help us connect the properties of neural models that we presented above with more general facts about mental representation and computation. They provide the symbolic backbone from which these models will get a lot of their interpretability. To embed them into the theory of causal explanations that I present here requires going down to the neural level however, which I propose to do in **Section 4.3**. This is largely facilitated by the central role that logic programs take in the field of *neuro-symbolic computation* (see for example Garcez, Broda, and Gabbay 2002; Garcez, Lamb, and Gabbay 2009), which seeks to integrate symbolic knowledge representation with neural network learning algorithms. I focus on one particularly relevant result in this line of research, which is the *Connectionist Inductive Learning and Logic Programming* (CILP) translation algorithm. CILP translation provides a way to translate a logic program LPinto a corresponding neural network N_{LP}. Every atomic variable in LPhas a corresponding neuron in N_{LP}. Each clause in $\mathsf{LP}$is mapped onto a small sub-network of neurons, whose connectivity and activation thresholds reflect the structure and logical dependencies of the source program. Effectively, it translates the program instantiated by clauses (4.4–4.5) into the network represented in Figure 4.2. It does not just provide the network structure represented in the figure, but also populates it with connection weights as well as biases for individual neurons. These weight parameters feed into $tanh(\dot{})$ activation functions for the hidden and output nodes, yielding activation values in $[-1, 1]$ — negative activations provide advantages for representing negations. Standard neural processes such as activation propagation can then approximate the program's truth-functional behavior via continuous functions.

The translation algorithm sets all weights parameters relative to a threshold A_{min} in a way that provably guarantees that each neuron's activation is always above A_{min} whenever the value of its inputs is such that the proposition represented by that program could be proven in the source program, and below $-A_{min}$ otherwise. This means that A_{min} can in principle be used as a linear threshold for neurons, to have the network encode the program's truth-functional behavior *exactly*. The mapping $\mathsf{LP}$to $\mathsf{N}_{\mathsf{LP}}$is direct and systematic. It depends on the structure of the source program only, and does not involve training. Because its hidden layer structure closely mirrors that of the source program, we can define operations directly over the network. These will account for the patterns of judgment that we traced back earlier on to the modular structure of logic programs, as well other facts that illustrate the added explanatory purchase of such network translation.

The first type of operation, which I explore in **Section 4.4**, is a *sampling procedure* over the networks generated in this way. The sampling procedure I have in mind is the following.

The network N_{LP}, which corresponds to subjects' internal representation of the causal system, is initialized in a state that corresponds to the *actual world* of reference, which they just observed or have been told about and about which we ask them to produce causal explanations. This means that the input nodes' activations

are set to values that track the occurrences of events in the real world (1 if an event happened, -1 otherwise), the value of all other events being then determined by activation propagation to the next layers. From this starting point, the subject explores neighboring states of the network by moving to an alternative state — corresponding to an alternative valuation over the input nodes — and then from that alternative state to a new one, and so on for a *small* number of steps. This captures the process at the heart of counterfactual theories, by which reasoners consider alternative scenarios to produce causal explanations. The probability of moving to any alternative state at each step is determined by the state of the network at the previous one — and independent from its state at steps before that one. This makes the procedure a Markov process. More specifically, I propose to understand it in a way analogous to Davis and Rehder (2020)'s *Mutation Sampler*, a variant of the Metropolis-Hasting sampling algorithm where the state considered for transition at each sampling step differs by the present state by the value of only one binary variable. This Markov sampling procedure captures the intuition that people explore alternative scenarios incrementally, typically considering only minor deviations from the observed actual state at each step.

Considering such an alternative state for the input variables at each step, subjects engaged in a process of considering counterfactuals make an internal decision either to move to that alternative state or to stay with the present one, with a probability that depends on:

1. The normality associated with the input event in question, which I propose to capture in the form of a bias on each input node. This accounts for the fact that normal scenarios are more likely to be considered by subjects as they engage in counterfactual simulations.

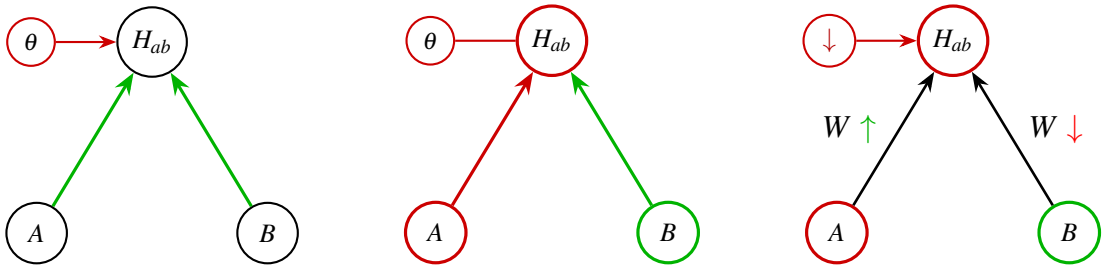
2. But also the value of the *hidden nodes* to which the input variable being considered is connected.

   Consider for illustration the subnetwork represented in Figure 4.3a, where the input nodes A and B are connected to the hidden node H_{ab}. Suppose that in the initial state, both A and B have activation value 1 and the weights on their connections to H_{ab} are positive, so that H_{ab}'s activation is also positive.

   Suppose then that the subject considers moving to the alternative state where $A = -1$. I assume that the probability that the subject actually undertakes this option will be influenced by the value of H_{ab}, in that a state of the input nodes is less likely to be accessed if it is going to change the activation value of the hidden nodes with it. In this example, the fact that having A go from $A = 1$ to $A = -1$ would change the value of H_{ab} from 1 to -1 once a forward pass is run on the network weights against changing the value of A, and in favor of sticking with the present state.

The fact that the sampling process is influenced by the activation values of hidden nodes, which are determined by the present state is important because it introduces *auto-correlation* in the sampling process. In other words, it introduces a tendency for each successive state to look like the previous, more so than they would have if sampling was determined by the normality of events alone. Assuming as we do that people only run such sampling process for a small number of steps, this in turn means that the alternative scenarios that they consider will tend to be very similar to the actual-world of reference. This tendency is desired, because it captures the long-standing assumption that counterfactuals tend to be sampled *close to the real-world* (Lewis 1973a; Lucas and Kemp 2015).

It also offers a partial explanation of some of the ways in which similarity to the real-world is itself structured by groupings of variables. For someone looking at a room filled with 5 red chairs, it seems easier to imagine an alternative room with 5 blue chairs, than one with 3 red chairs and 2 blue ones, in spite of the fact that the latter is closer to the real-world in some sense. Similarly, in the example given above, once a subject has moved to the alternative state where $A = -1, B = 1$ (as in fig. 4.3b) in spite of the reluctance previously introduced by $H_{ab} = 1$, in this new alternative state the value of H_{ab} switches to $H_{ab} = -1$. As a



(a) Subnetwork (b) Sampled state (c) Weight updates

Figure 4.3: An illustration of the network fragment for A, B, and H_{ab}. a: Subnetwork fresh out of translation. b: Sampled state: A, H_{ab} inactive (red), B active (green). c: Weight updates: reward $A \rightarrow H_{ab}$ ($\uparrow$ in green) and penalize $B \rightarrow H_{ab}$ ($\downarrow$ in red).

result, further transition to a state where, e.g. $A = -1, B = -1$ encounters fewer obstacles. I will illustrate how this feature of the sampling process captures some intuitive properties of the ways in which we explore alternative scenarios.

In **Section 4.5** I show how this normality-sensitive sampling process will translate into the patterns of causal responsibility judgments, by means of a process of weight updates in each of the network states visited. I present a weight update procedure known as Layer-Wise Feedback Propagation (LFP) (Weber et al. 2025), inspired from the Layer-Wise Relevance Propagation procedure for explaining deep neural networks predictions (Bach et al. 2015; Montavon et al. 2017; Montavon, Samek, and Müller 2018). It has the advantage of being less sensitive to vanishing gradient problems that make it hard to use procedures like gradient backpropagation (Rumelhart, Hinton, and Williams 1986) on networks with already "mature" weights like the ones we're dealing with. Its logic is the following:

1. Assign a certain "reward score" to the outcome node (E in our example) of the network.

2. Propagate that reward to nodes in next (i.e. hidden) layer, as a function of how much each contributed to giving that node the activation value that it has in a certain state. If the outcome node is positively activated, for example, the hidden nodes that send positive input to it are rewarded. As a result the *weights* on the edges between those and the outcome are *increased*, while those coming from nodes that send negative inputs are *decreased*.

3. Apply the same process to the connections between hidden and input nodes.

The situation depicted in fig. 4.3b–c illustrates how such a process is going to capture the pattern of *abnormal inflation*. Suppose that event A is seen by a subject as more normal than event B. Then, by construction, that subject will sample more states where $A = -1$ (and thus $H_{ab} = -1$) while $B = 1$ (as in fig. 4.3b) than the reverse. In such states, A's contribution is aligned with H_{ab}'s (because A sends negative-valued input and H_{ab} has itself a negative activation value) whereas B's contribution is misaligned. As a result, we reward A's connection by increasing $w_{A \rightarrow H_{ab}}$ and penalize B's connection by decreasing $w_{A \rightarrow H_{ab}}$ (as pictured in fig. 4.3c).

A subject might also consider states where both A and B are active or inactive. In those cases however, both connection weights will be rewarded, so no contrast is introduced between the two as a result of visiting such states. As I will explain a few paragraphs below, such contrasts are ultimately what matters for causal importance. Since those are fully determined by the relative proportion of states $(A = -1, B = 1)$ compared

to $(A = 1, B = -1)$, which is itself determined by the relative normality of A and B, we expect abnormal inflation to emerge at the level of input-to-hidden connection weights.

One can then use a similar logic to reconstruct how *abnormal deflation* is expected to emerge with respect to hidden-to-output nodes. The disjunctive-gate nature of hidden-to-output connections means the output node is always active in "contrast cases" (where some hidden nodes are active but others aren't). This in turn favors connections from the subset of nodes that *are* active, hence ultimately from those nodes that are more *often* active due to normality.

This process of weight updates sets the basis that will then allow us to "read" causal importance scores directly off the weight parameters of an internal model (in ways that I describe in the presentation of section 4.6 a few paragraphs below), recapturing an insight present in associationist theories of causal strength mentioned in the introduction to this dissertation. The explanatory purchase afforded by a theory that maps causal importance scores directly onto parameters of an internal generative model will be most conspicuous later in Chapter 7 of this dissertation, when we look at how causal explanations help subjects *learn* causal rules from observations.

For the present context, it may be worth remarking how a theory of this shape puts us in a good position to explain some intuitively plausible features of our explanations (though hard evidence for them would be difficult to construct) that are hard to fit in an account that sees it purely in terms of covariation across counterfactuals. One is the possibility that our judgments may in part be shaped by observations directly in some instances (repeated co-occurrence of events might influence our intuitions of causal importance by tuning weight parameters via error-driven learning processes). Another is the introspective intuition that the insights that we gain from reflecting about events "stay" with us even after we've stopped thinking about them, in the form of emphasis put on certain key variables. The long reflection by which I consider possible explanations for why I did not wake up on time for some important meeting (considering the importance of various factors such as my forgetting to set an alarm, jetlag, a general seasonal fatigue by varying counterfactuals) can lead me to emphasize some of those factors in such a way that I'll spontaneously be driven to pay more attention to them on future occasions (e.g. it makes it harder to forget to set my alarm again). And having gone through such a reflective process intuitively seems to yield more reliable "conditioning" than that brought about by the unreflective displeasure of the original experience alone.

Together, such features would help make sense among other things of the intuition that the presence of oxygen in the air is not merely an unimportant cause of forest fires "when we think about it" by contemplating counterfactuals, but precisely because we *don't even think about it* to begin with, owing to the fact that our past reflective and direct experience conspire to lead us to see it as immaterial.

Finally, **Section 4.6** explains how the weights resulting from the update process will be interpreted as measures of causal importance. The measure will involve two components:

1. A process for assigning *relevance* to input variables, which rests on the the Layer-wise Relevance Propagation procedure (Bach et al. 2015; Montavon et al. 2017; Montavon, Samek, and Müller 2018) originally developed to explain the predictions of deep neural networks and briefly mentioned above. Its principle is as follows. After having updated the weights on one's inner model via the sampling-and-update process described above, one goes back to the initial state (representing events as they happened in the real world) and then "trickles down" a relevance score from the output node at the top through the hidden layer and all the way down to the input nodes at the bottom. As relevance signals travel through the network they spread through each branch proportionally to the relative contribution of nodes in lower layers to the activation input of the nodes in upper layer from which they inherit relevance. Such contributions are directly proportional to the weights on the respective connections, which ultimately means that an input cause gets all the more relevance when it is more strongly connected to nodes

in the hidden layer, and that those hidden nodes to which it connects are themselves more strongly connected to the relevant outputs.

LRP-measures possess certain degrees of freedom that allow us to, for example, have relevance signals travel only through positive contribution paths, which we will assume people do in the cases we look at. This comes with a great theoretical advantage, as doing so dispenses with having to assume a "pre-selection" step by which the relevant set of causal candidates for causal selection is defined on the basis of theories of actual causation (**halpern:2005**; like e.g. Halpern 2015). Current theories of causal selection all (explicitly or implicitly) assume such a step as a precondition to computing the causal importance of the chosen candidates. As we will see, using a LRP-measure in which signals only travel through positive paths (or negative paths in the case of negative outcomes) automatically results in assigning zero relevance to those variables that aren't actual causes of the outcome in the context of reference. This provides a path to unify theories of actual causation with theories of graded causal importance through a single measure.

2. A measure of *path complexity* that allows us to convert the relevance for a set of causal variables C into a measure of causal importance $\kappa(C,O)$ by dividing the sum of the relevance held by causes in C by the number of edge-disjoint routes by which these causes contribute to the outcome:

$$\kappa(C,O) \;=\; \frac{\sum_{c\in C} R_c}{\mathscr{C}(C,O)}, \tag{4.6}$$

where $\sum_{c\in C} R_c$ is the total relevance of the causes in C, and $\mathscr{C}(C,O)$ is a factor tracking the number of *active routes* from C to O (in a sense essentially similar to Hitchcock 2001a) but counting only the number of edge-disjoint paths. The $\mathscr{C}(C,O)$ denominator is especially relevant for plural cause judgments (where C contains more than one variable). It implements the reluctance for explanations that borrow from disjoint sets of sufficient conditions for satisfying the outcome, described earlier in terms of using several clauses of a logic program.

It will also allow us to gracefully handle the reversal observed for negative conditions. Following the logic according to which the network originating from the logic program $\{E \leftarrow A,B; \quad E \leftarrow C,D\}$ only allows us to derive $\sim E$ (failure to derive E), we assume that to explain the losing outcome $\neg E$, people have to filter relevance through an additional path $\neg E \longleftarrow\!\sim E$ before it can trickle down through the rest of the network. This in turn means that there are no two edge-disjoint paths from inputs to the explanandum $\neg E$ (since all paths go through $\neg E \longleftarrow\!\sim E$), which makes the $\mathscr{C}(C,O)$ equal to 1 for all causes singular or plural, thus encouraging prolixity.

Sections 4.2 to 4.6 focus on theoretical presentation. I also operationalize the theory into a model coded in `R` (and available on the public repository at https://osf.io/zv2m5/), and show such a model allows us to capture the response patterns collected in the experiment on plural causes presented in Chapter 3 much better than existing models while fitting fewer parameters to the data. This neural approach is significant not only because it accounts for empirical data more effectively, but also because it integrates symbolic insights from logic-based theories of representation with empirically motivated computational constraints from neural network theory. I also coded a Rshiny interactive App, where readers can play with the different component of the model to see their interaction. Readers can access it at the address: https://konuk-can.shinyapps.io/myshinyapp/

4.2 Logic Programs

Logic programming in the general sense refers to uses of logic in computer programming. The related, but more specific notion of logic programming we have in mind here refers to a general declarative programming paradigm. It does not constitute by itself a concrete, computable implementation; rather, it serves as a common semantic foundation for logic-based reasoning in computer science. Its advantage here is that it provides us with an *abstract framework* for thinking about the way in which logical relations can be implemented in concrete programs, including as a special case the internal (mental) programs that underly human subjects' grasp of those relations. This abstraction makes it ideally suited for establishing connections between such programs and independently-known features of mental representations.

This is what I do in this section. I first offer a brief presentation of the class of Definite logic programs in section 4.2.1. This is the simplest class of programs, which only deals with relations that can be represented using positive literals only. This enables me to highlight several general features of the way in which propositions can be proven in logic programs, which I connect with independent observations about how people grasp the meaning of propositions in mental model theories of representation; in particular, I emphasize how the proof procedures of logic programs naturally focus on minimal sufficient conditions for proving a certain proposition. Then, in section 4.2.2 I present the more expressive class of General logic programs, which includes a representation of negated literals $\sim L$, where $\sim L$ is understood as *failure* to derive L. Negation as failure differs from the classical negation $\neg E$, and is also understood to be less demanding in the framework of logic programs, in the sense that handling classical negation requires considering a larger class than general logic programs (known as *extended* logic programs). This gives a suggestion that negation should by default be handled in terms of negation as failures, which I argue is also an appropriate assumption to make about the way humans handle negative outcomes. This assumption will have several consequences later on our understanding of how humans handle negative outcomes.

4.2.1 Definite logic programs and SLD-resolution

Definition 1 (Definite logic programs)**.** *A **definite logic program** is a finite set of clauses of the shape*

$$A \leftarrow B_1, \ldots, B_n$$

where A and each B_i are propositional atoms.

The single atom A constitutes the *head* of the clause, while $B_1, \ldots, B_n$ together constitute its *body*. One can distinguish two kinds of clauses in logic programs in general: *Rules* and *Facts*. A *rule* is a clause with a non-empty body, while a *fact* is a clause with an empty body, like:

$$F \leftarrow$$

which can also be denoted as:

$$F \leftarrow \top, \quad \text{or simply } F$$

Intuitively, such clauses are a logic program's way of representing that F is true. Proving propositions in logic programs will then amount to combining clauses to turn rules into facts. For definite logic programs, this is done using a procedure known as **SLD-resolution** (Selective Linear Definite clause resolution), presented below.

SLD-Resolution

Definitions

- **Goal (Query):** A finite set of literals $G = \{L_1, L_2, \ldots, L_m\}$ representing what we wish to prove or compute.
- **Derivation:** A sequence of goals $G_0, G_1, G_2, \ldots$ where each goal G_{i+1} is derived from G_i by applying a resolution step.
- **Resolution Step:** The process of selecting a literal in the current goal and resolving it with a clause from the program.

SLD-Resolution Procedure

1. **Initialization:** Start with the initial goal G_0.
2. **Selection Rule:** At each step, select a literal L from the current goal G_i.
3. **Resolution:**
   - Find a clause in the program $A \leftarrow B_1, \ldots, B_n$ such that $A = L$.
   - Replace L in G_i with the body literals $B_1, \ldots, B_n$ to form the new goal G_{i+1}, s.t.:

$$G_{i+1} = (G_i \setminus \{L\}) \cup \{B_1, \ldots, B_n\}$$

4. **Termination:**
   - If the goal G_{i+1} is empty, the derivation **succeeds** (the query is proved).
   - If no applicable clause can be found for the selected literal, the derivation **fails**.

Effectively, SLD-resolution proceeds by successively replacing each atom B_i in the body of a clause $A \leftarrow B_1, \ldots, B_n$ with the body of *another clause* in the same program that has B_i as head. Ideally, the body of that clause is empty (the clause is a *fact*),(a *fact*), immediately satisfying the condition B_i and allowing us to move on to satisfy other atoms, eventually turning A into a fact itself. But it might also involve going through intermediate steps whereby we replace B_i with some other atom(s) before we can satisfy those. Consider for example the following program:

$$E \leftarrow A, B$$
$$A \leftarrow$$
$$B \leftarrow C$$
$$C \leftarrow$$

And suppose we set $\{E\}$ as our initial goal (we want to prove E). The first clause tells us that the only way to do that is to replace E with the pair of subgoals A, B, resulting in the new goal $\{A, B\}$. We might then try to resolve each of A and B in succession. The resolution of A demonstrates the ideal case mentioned above: since $A \leftarrow$ is a fact (empty body), we can immediately satisfy this subgoal, reducing our focus to $\{B\}$. To prove B, however, requires an intermediate step through C, since the program does not include a fact $B \leftarrow$ but only offers to satisfy B via $B \leftarrow C$. This substitution creates a new subgoal $\{C\}$, which can in turn be satisfied via $C \leftarrow$.

With no remaining subgoals ($\emptyset$), the derivation succeeds and we can take the original goal E (as well as every intermediate subgoal) to be proven. This sequence exemplifies both direct satisfaction through facts (A and C) and the need for intermediate substitutions ($B \rightarrow C$), illustrating how SLD-resolution integrates

immediate solutions with stepwise decomposition of complex goals. One can also see through this example all of the ways in which the derivation of E could have *failed*, if any of the three other clauses were absent.

The proof procedure instantiated by such programs is still very abstract. It leaves out such details as how each propositional atom is represented in an inference engine, the order in which one iterates over subgoals, or over clauses that share the same head. Still, it highlights some of the additional constraints entailed by the necessity to make truth functions executable.

The closed-world assumption in logic programs. First, the necessity of resolving atoms by other clauses in the program highlights how derivations are sensitive to those relations an agent *cares to represent* (and therefore compute), rather than all relationships that may generally hold. This is a hallmark of *nonmonotonic* inference, a concept widely recognized to be relevant in human reasoning as well. A well-known illustration comes from the *suppression effect* via alternative causes (Byrne 1989). Suppose a reasoner is presented with the pair of premises:

(1) If she has an essay to write, then she will study late in the library.
(2) She has an essay to write.

Most people naturally conclude that she will study late in the library, correctly endorsing the inference from (1) and (2). However, people also commonly endorse certain reasoning fallacies, notably:

Affirming of the Consequent (AC): From "she will study late in the library" people tend to infer "she has an essay to write."

Denying the Antecedent (DA): From "she does *not* have an essay to write" infer "she will *not* study late in the library."

Although these patterns are invalid in classical logic, their frequent acceptance by participants can be understood from a *logic program* perspective, combined with a *closed-world assumption*(Stenning and Lambalgen 2008). To illustrate the phenomenon, consider that a reasoner represents premises (1) and (2) by the following minimal program:

$$\text{library} \leftarrow \text{essay},$$
$$\text{essay} \leftarrow$$

Where the first clause reads "`library` is true if `essay` is true," while the second is a fact stating "`essay` is true." Applying SLD-resolution to prove `library` from these clauses is straightforward: one uses the first clause to reduce `library` to `essay`, and the second clause (a fact) then resolves `essay` with the empty body, thus proving `library`. However, under the closed-world assumption, an agent also assumes that all relevant clauses that could prove `essay` or `library` are in the program. Hence, since the only derivation to satisfy `library` involves satisfying `essay` as one of its subgoals, it becomes natural, under the closed-world assumption, to infer from the truth of `library` that `essay` must also be true, thereby producing the fallacious AC inference.

Likewise, telling reasoners that they can't derive `library` from `essay` (by telling them that essay is false), would lead one to assume that `library` is false under the closed-world assumption, yielding the DA fallacy. Support for this closed-world interpretation comes from the *suppression effect*, triggered when reasoners are offered a new premise, for instance:

(3) If she has a textbook to read, then she will study late in the library.

Now, people still typically preserve the valid inference from (1) and (2) — that having an essay to write implies she will study late. But they often "suppress" the previous fallacies (AC and DA). From a logic programming viewpoint, we can model this by enriching the program with a third clause:

$$\begin{aligned} \texttt{library} &\leftarrow \texttt{essay}, \\ \texttt{library} &\leftarrow \texttt{textbook}, \\ \texttt{essay} &\leftarrow \end{aligned}$$

Now there is a second way (`textbook`) to satisfy `library` besides `essay`. As a result, reasoners are made aware of the fact that there is a potential derivation for `library` that does not involve `essay` as a subgoal. Similarly, blocking the path library $\leftarrow$ essay (by "denying the antecedent" `essay`) does not invite the conclusion that there are no routes left to derive `library`.

The modular nature of logic programs and the emphasis on alternatives. Another significant feature of logic program semantics is how individual clauses highlight alternative pathways to derive a particular outcome. This relates to a striking aspect of the semantics of logic programs, which is the fact that they do not allow for classical disjunctions to be introduced in the body (or indeed in the head) of clauses. This is true even for more expressive classes of programs, as shown in section 4.2.2. The list of literals in the body of a clause is always to be interpreted conjunctively as a list of subgoals, *all* of which need to be satisfied in order to derive the head. This means that functions which in structural causal models can be represented as a single equation, such as the rule:

$$E := (A \wedge B) \vee (C \wedge D) \tag{4.7}$$

would have to be represented in terms of two separate clauses in a logic program:

$$E \leftarrow A, B \tag{4.8}$$
$$E \leftarrow C, D \tag{4.9}$$

This property makes sense in the context of the SLD-resolution procedure. Allowing for a disjunctive clause like

$$E \longleftarrow (A \wedge B) \vee (C \wedge D)$$

would suggest that the SLD-resolution procedure should branch out into two parallel search process, one trying to resolve $\{A, B\}$, the other $\{C, D\}$, thereby increasing the complexity of the procedure. By contrast, by encapsulating each disjunct into a separate clause, we *modularize* the resolution procedure, ensuring that it only goes through both clauses in the *worst case*, when the procedure fails to prove E through the first one it tries out.

This modular aspect of logic programs also mirrors independent observations made in theories of mental representations. In particular, Koralus and Mascarenhas (2013, 2018) show how many patterns of inferences in reasoning can be captured by assuming that humans treat disjunctive propositions as *questions*, where a question is represented as a set of alternative models (like $\{a \wedge b,\ c \wedge d\}$), and then try to answer those by using procedures that are inherently sensitive to the structure of the set of alternatives in question. These procedures capture consistent patterns of fallacious inferences such as *illusory inferences from disjunction* (Walsh and Johnson-Laird 2004) from series of premises like the following:

P1: *Alice speaks French and Chinese, or she speaks English.*

P2: *Alice speaks French.*

Q: Does it follow that Alice speaks Chinese?

When presented with a problem of this shape, from P1 and the fact given in P2 that Alices speaks French, subjects typically infer that Alice speaks Chinese as well. This highlights the way in which people's representation of the disjunction in P1 does not match the classical disjunction $(\texttt{french} \wedge \texttt{chinese}) \vee \texttt{english}$, because the conjunction of that conjunction with the proposition `french` affirmed by P2 would not logically entail `chinese`, because premise P1 remains true even if Alice does not speak Chinese, provided she speaks English instead. Koralus and Mascarenhas (2013, 2018) propose to understand the fallacy by assuming that the first premise raises a set of alternative models $\{\texttt{french} \wedge \texttt{chinese}, \texttt{english}\}$, and that reasoners subsequently use procedures to *answer the question* on the basis of the second premise `french`. These lead them to retain only the alternative that contains `french` (i.e. $\texttt{french} \wedge \texttt{chinese}$) yielding the fallacy.

Although the procedures that Koralus and Mascarenhas have in mind are not the same as the SLD-resolution procedure defined above, the two have in common that they are sensitive to a notion of *alternatives* (implicitly in the case of logic programs, in the form of alternative resolution paths) that does not reduce to classical disjunction.

4.2.2 Negation as failure in logic programs

In the previous subsection, I explained how the clause structure of logic programs helps to simplify computation in a goal-oriented proof procedure. It allows one to prove a certain proposition without having to deploy the entirety of one's logical knowledge. In this section, I show an extension for the treatment of negations. Proving the classical negation $\neg P$ of a proposition is typically harder than proving P because it involves making sure that no set of rules compatible with our knowledge would allow us to prove P.

To cope with this issue, logic programs typically use a more cautious notion of negation known as *negation-as-failure* or also *default negation* (noted $\sim P$), which allows one to relativize negation to a particular logic program via the closed-world assumption. It enables deriving $\sim P$ from a *failure to derive* P using the clauses of the program. Unlike classical negation, negation-as-failure ($\sim P$) does not assert that P is false across all logical models; rather, it expresses the weaker claim that P is not provable given the current knowledge base.

Negation as failure is introduce by extending the class of definite logic programs defined above to allow for *general clauses* (also called *normal* clauses) that include literals with default negation in the body of the clause.

Definition 2 (General Clause)**.** *A* general clause *is a rule of the form*

$$A_0 \leftarrow A_1, \ldots, A_m, \sim A_{m+1}, \ldots, \sim A_n,$$

where A_i *(*$0 \leq i \leq n$*) is an* atom *and* $\sim$ *denotes* default negation*. A* general logic program *is a finite set of general clauses.*

The addition of general clauses makes logic programs highly expressive in spite of the absence of classical negation. In Chapter A I argue (with a proof sketch) that programs of this class can represent all of the relations that can be represented in deterministic SCM restricted to boolean functions. Readers interested in more detail about the semantics of SCMs can also find them there. The introduction of general clauses comes with an extension of the SLD-resolution procedure known as SLDNF-resolution (for Selective Linear

Definitive with Negation-as-Failure), presented below. It allows one to satisfy default-negated literals in the bodies of clauses from a failure to derive the corresponding positive literal:

SLDNF-Resolution Procedure

1. **Initialization:** Start with initial goal G_0
2. **Selection Rule:** Choose literal L from current goal G_i, then:
   - **If L is a positive literal:**
     - Apply **SLD-Resolution** as described above.
   - **If L is a negated literal $\sim A$:**
     - Attempt to derive A using SLD-resolution.
     - If the derivation of A **succeeds**, then $\sim A$ **fails**.
     - If the derivation of A **fails** ($\sim A$ succeeds), **remove** $\sim A$ **from** G_i to form G_{i+1}.

To illustrate this process, consider the following program:

$$A \leftarrow B, \sim C$$
$$B \leftarrow$$
$$C \leftarrow D$$

And suppose we want to prove A. The derivation begins with the initial goal $\{A\}$. Using the first clause, we replace A with the subgoals B and $\sim C$, resulting in the new goal $\{B, \sim C\}$. The subgoal B can be immediately satisfied via the second clause using SLD-resolution, leaving us with $\{\sim C\}$. To satisfy $\sim C$, following SLDNF-resolution, we then attempt to satisfy C:

- We look for a clause with C as head. Only one clause ($C \leftarrow D$) can be found, leading us to the new goal $\{D\}$.
- We look for a clause with D as head, but none can be found. The goal cannot be satisfied so our derivation of C fails. Hence by SLDNF, $\sim C$ succeeds.
- With both B and $\sim C$ satisfied, the derivation of A succeeds with empty goal $\emptyset$. So our original goal A is proved.

The introduction of negation-as-failure makes general logic programs very expressive (on that topic see Appendix Chapter A where I offer to translate every boolean SCM into a general logic program in a way that preserves essential truth-relations) without introducing a great deal of additional machinery compared to basic Definite programs, by leveraging the closed-world assumption. In particular, general logic programs do not allow derivation of explicitly negated literals $\neg L$. This makes the framework of general logic appropriate for capturing, the notion central to mental models type of accounts that the way in which we internally represents events and the relationships between focuses on the events that effectively happened, not on the events that didn't happen. This notion becomes clearer when considering mental simulations we run about a domain of events, a notion which counterfactual theories recognize as central to our understanding of causal relations. One always simulates *occurrences* of events; in fact it is not even clear what it would mean for someone to simulate the *non-occurrence* of an event. At best one can represent that running one's internal simulation

model with certain input settings *does not generate an occurrence* of a certain event E. Which is exactly the notion of negation-as-failure $\sim E$ that general logic programs capture.

For this reason I am going to assume in general that by default people will represent the causal domain described by a rule like (4.7) using an internal program whose behavior is captured, at a high-level of idealization, by rules like those in (4.8)–(4.9) for generating occurrences of E. A limit implicit in this framing is that this program does not have a mechanism for generating occurrences of $\neg E$. This will be a problem when people are explicitly asked to explain why a negative outcome $\neg E$ happened, as in the experiment presented in Chapter 3 when we ask people to explain why a player *lost*. This is because the program instantiated by rules (4.8)–(4.9) only gives them a way to tell that, given certain ground facts (certain draws from urns) the round is not won. To go from there to the knowledge that one lost may be trivial, especially given that the subject can safely make the closed world assumption, i.e. assume that their model represents all the relevant facts for determining the outcome of the round. Crucially however, that assumption is not baked in by default in a general program that can't represent classical negation, but has to be explicitly. Much later in this chapter, I will propose to understand it in terms of a dedicated closed-world operator that explicitly encodes the transition $\neg E \leftarrow\sim E$. Before I can introduce it and explain how it accounts for how people handle the explanation of negative outcomes however, I need to introduce additional machinery and in particular explain how the program features that I have here described at a high, symbolic level of idealization are instantiated in neural models at a lower level. To understand precisely how these symbolic structures and assumptions manifest in neural architectures, the next sections will systematically translate these symbolic insights into concrete neural implementations.

4.3 The CILP translation algorithm (Garcez, Lamb, and Gabbay 2009)

4.3.1 Intermediate summary and presentation of the section.

The preceding section introduced logic programs as a symbolic framework for modeling mental representations. I highlighted their capacity to modularize proof pathways and handle negation through goal-directed procedures. I emphasized the alignment between properties of these programs and independently described psychological phenomena, such as non-monotonic reasoning patterns and the treatment of disjunctions as alternatives. These properties make logic programs a useful descriptive framework, at a symbolic level of idealization, for how people internally represent the causal knowledge that they leverage as they elaborate causal explanations — at least for the sort of causal relations that we usually describe in terms of boolean functions. In particular, the class of general logic programs, which allows default negation in the bodies of clauses, can in principle serve as complete framework for representing all of these relations. This framework would arguably be inadequate in some cases where the use of classical negation would be required (see, e.g. Gelfond and Lifschitz 1991) but we will allow ourselves to ignore them for the present presentation.

The framework of logic programs plays for us the role of an intermediate level of idealization, which highlights elements of the structure of the internal programs that subjects use for representing causal relations which the descriptive angle adopted by Structural causal models tends to obviate. Ultimately however, I contend that the graded nature of causal judgments is best understood by a connectionist view that uses these programs as a basis to formulate hypotheses about neural models with continuous dynamics. This is what I propose to do in this section, where I show how the same relations captured at a symbolic level by general logic programs can be translated rigorously into neural models whose dynamics reproduce the

program's truth-functional behavior via continuous functions. My task is greatly facilitated by the fact that the relationship between logic programs and neural models has already been abundantly explored in the context of research on neuro-symbolic architectures. The section is dedicated to introducing one product of that research, the Connectionist Inductive Learning and Logic Programming (CILP) algorithm (Garcez, Broda, and Gabbay 2002; Garcez, Lamb, and Gabbay 2009), a translation algorithm from general logic programs into neural networks.

My presentation here follows the version proposed in Garcez, Lamb, and Gabbay (2009) and does not add any new elements. The neural models resulting from that translation will provide a substrate for the processes of probabilistic sampling, weight updates, and relevance propagation that I define in later sections.

4.3.2 Translating Logic Programs into Neural Networks: The CILP Algorithm

The CILP algorithm (Garcez, Broda, and Gabbay 2002; Garcez, Lamb, and Gabbay 2009) provides a translation algorithm from general logic programs, as defined in the previous section, into equivalent network representations.

Given a general logic program LP, comprising a set of general clauses C, CILP constructs a neural network N_{LP} with the following properties.

Layer structure. N_{LP} contains three layers: an input layer, a hidden layer and an output layer. The structure of each layer is fully determined by the clauses in the source logic program LP:

1. The **Input layer** contains one neuron per each unique atom appearing in the *body* of any of the clauses $C \in \mathsf{LP}$.
2. The **Hidden layer** contains one neuron per clause $C_i \in \mathsf{LP}$. These encode the program's logical pathways.
3. The **Output layer** contains one neuron per unique atom appearing in the *head* of any of the clauses $C \in \mathsf{LP}$.

When the source program includes atoms that appear both in the head of one clause and in the body of another (e.g. $\{A \leftarrow B, B \leftarrow C\}$, these are represented by two distinct neurons representing that atom, one in each layer, and drawing a recurrent link between the neuron in the output layer and that in the input layer. We will not examine such recurrent structures here however, so they can be ignored for the current purposes.

Activation values and functions. Neurons in N_{LP} take values in the interval $[-1,1]$. For hidden and output neurons, their activation function is the hyperbolic tangent

$$a_n = \tanh(x_n), \quad \text{with } x_n \text{ the net input to a neuron } n \tag{4.10}$$

The hyperbolic tangent is a non-linear activation function. It is much like the better known sigmoid function in that it smoothly squashes any real number input into a bounded range, and has the same characteristic s-shaped curve. It differs in that that its codomain ranges from $[-1,1]$ and it centers on zero[1]. The activation

[1] More precisely, the CILP algorithm allows any version of the *bipolar activation function* $f(x) = \frac{2}{1+e^{-\beta x}} - 1$ as the activation where $\beta > 0$ controls the steepness of the transition between -1 and 1; $\tanh()$ is the special case of that function where $\beta = 2$. We make this simplification because, in the CILP framework, the parameter β is absorbed into the definition of the weight parameter W (presented below) via the product $\frac{2}{\beta}$. The choice of $\beta = 2$ allows us to omit β from both equations by making $\frac{\beta}{2} = 1$.

function for input neurons is the standard linear activation function

$$a_n = x_n, \quad \text{with } x_n \text{ the net input to a neuron } n \tag{4.11}$$

Essentially, this allows the input activations to be set by any input vector I provided to the network, for example during training. For our purposes, the input vectors will be determined by the sampling procedure that we define in the next section.

Connectivity patterns and parametrization. Each hidden neuron H_i receives connections from all literals in the body of clause C_i. Each output neuron receives connections from all hidden neurons H_j whose clause heads match the output atom. The weights on **all** edges of the network share the same absolute value W, determined as a function of two parameters:

- $\text{MAX}_{\mathsf{LP}}(\vec{k}, \vec{\mu})$, defined as the maximum of the following two quantities:
  1. the highest number of literals in the body of any given clause $C \in \mathsf{LP}$,
  2. the highest number of clauses sharing the same head in LP.
- The threshold value $A_{\min} \in (0,1)$. The weight parametrization of the CILP algorithm depends on that threshold in such a way that provably ensures (see Garcez, Lamb, and Gabbay (2009)'s **Theorem 8**) that the activation of each neuron in N_{LP} is always above A_{min} whenever the corresponding atom can be proven in the source program LP by taking as facts the same input atoms whose activation is set to 1 in N_{LP}, and below $-A_{min}$ otherwise. The algorithm allows some freedom to set that value, but it is bounded as a function of $\text{MAX}_{\mathsf{LP}}(\vec{k}, \vec{\mu})$ via:

$$A_{\min} > \frac{\text{MAX}_{\mathsf{LP}}(\vec{k}, \vec{\mu}) - 1}{\text{MAX}_{\mathsf{LP}}(\vec{k}, \vec{\mu}) + 1}. \tag{4.12}$$

For the present purposes and to ensure modeling stability and reproducibility, the code accompanying this dissertation stipulates that A_{min} is always equal to

$$\frac{\text{MAX}_{\mathsf{LP}}(\vec{k}, \vec{\mu}) - 1}{\text{MAX}_{\mathsf{LP}}(\vec{k}, \vec{\mu}) + 1} + \frac{1}{\text{MAX}_{\mathsf{LP}}(\vec{k}, \vec{\mu})},$$

which ensures satisfaction of the inequality above (4.12) while also staying in the $(0,1)$ interval. The weight parameter W is then bounded by:

$$W \geq \frac{\ln(1 + A_{\min}) - \ln(1 - A_{\min})}{\text{MAX}_{\mathsf{LP}}(\vec{k}, \vec{\mu})(A_{\min} - 1) + A_{\min} + 1}. \tag{4.13}$$

For simplicity and again to ensure stability, my implementation also assumes that W is always exactly equal to that bound. The input-to-hidden connection weights are set to W wherever the input of the corresponding clause is a positive literal, and to $-W$ whenever it is a negated literal $\sim L$. Finally, the threshold θ_{H_i} of each neuron H_i in the hidden layer is set to:

$$\theta_{H_i} = \frac{(1 + A_{\min})(k_i - 1)}{2W}, \text{with } k_i \text{ the number of inputs to } H_i \tag{4.14}$$

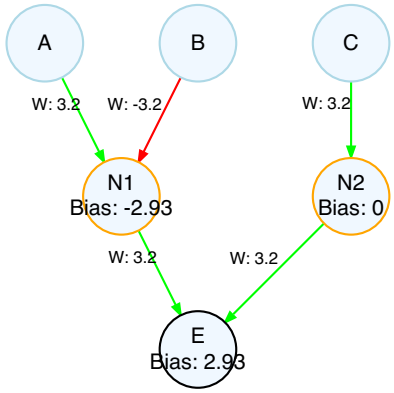


Figure 4.4: The Neural Network N_{LP_1}. Recall that neurons take activation values in $[-1,1]$. The values of the weights on this network make it clear that (1) the hidden neuron $N1$ gets a positive net input — and therefore a positive activation value close to 1 via $\tanh()$ — everytime $A = 1$ and $B = -1$, and a negative net input otherwise; (2) the hidden neuron $N2$ gets a positive activation value every time $C = 1$, and a negative activation value otherwise; (3) the output neuron E gets a positive activation value every time either one of $N1, N2$ is active, and a negative activation value otherwise.

And the threshold θ_{O_i} of each neuron O_i in the output layer is set to:

$$\theta_{O_i} = \frac{(1+A_{\min})(1-\mu_i)}{2W}, \text{with } \mu_i \text{ the number of inputs to } O_i. \tag{4.15}$$

This threshold setting ensures that the relations between activations at the level of input-to-hidden connections approximates conjunctive gates, and similarly approximates disjunctive gates at the level of hidden-to-output connections.

4.3.3 Example

As an illustration, suppose we want to translate the general logic program LP_1 below:

$$\mathsf{LP}_1 = \{E \leftarrow A, \sim B; \quad E \leftarrow C\}$$

Without any additional parameters, running the CILP translation algorithm as defined above on LP_1 yields the neural network N_{LP_1} in Figure 4.4.

The relations between activation values in the network that emerge from the weight values set by the equations above are in correspondence with the inference relations instantiated by the source logic program at a symbolic level. This correspondence is *approximative* if one merely runs the activation functions over hidden and output layers (via a forward pass) after setting input activations, meaning that atoms that are provable (given the chosen input facts) in the source programs would have activation values close to 1, while atoms that are not provable would have activation values very close to -1. It can also be made *exact* by using the $A_{\min}$ parameter defined above as a linear threshold, which is equivalent to passing the output a_n of $\tanh()$

into a step function $s(a_n)$ such that

$$s(a_n) = \begin{cases} 1 & \text{if } a_n \geq A_{\min} \\ -1 & \text{if } a_n \leq -A_{\min}. \end{cases}$$

The CILP equations guarantee that activation values of the network will always meet one of these two conditions as long as the input activations are in $\{-1, 1\}$. This discretization puts the continuous approximation into an exact binary correspondence, which exactly mirrors the true/false semantics of symbolic logic. Note, however, that the network could theoretically access activation values in the interval $(-A_{\min}, A_{\min})$ (e.g., by accepting input values not in $\{-1, 1\}$ or through adjustments to the network's weights). This intermediate range opens intriguing modeling opportunities for representing *partial* or *unknown* truth values in the style of fuzzy logic frameworks, but we don't look into them in this dissertation.

The translation provided by the CILP algorithm is foundational: it offers a precise bridge from discrete symbolic logic programs to continuous neural networks. It provides a framework for understanding how discrete logical rules might ground into neural representations with continuous weights. The next sections will build on those representations to explain how people can harbor *graded* and contrasted intuitions of causal importance for variables whose respective contribution they are simultaneously able to recognize as exactly symmetric on some other level — as when they recognize that A and B are equally necessary for E in $E \leftarrow A \wedge B$.

4.4 A sampling procedure over neural networks

4.4.1 Intermediate summary and presentation of the section

The preceding sections established how logic programs—declarative representations of causal knowledge—can be systematically translated into neural networks via the CILP algorithm. This translation preserves the inference relations of the source programs while implementing them in a neural substrate capable of continuous computation. The resulting neural network is structurally isomorphic to the source program: the clause structure of the program is reflected in the hidden layer structure of the network that translates it.

This translation captures the insights from symbolic theories of mental representation within connectionist models that can perform continuous computation. By itself, the translation does not introduce any weight imbalances between the different input variables. The absolute value of the weight W is the same on all edges of the network coming out of translation. This is appropriate in a context (like that of many experiments on causal selection judgments including the one presented in Chapter 3) where the causal rule underlying a domain is framed in a way that makes the contribution of each variable exactly symmetric.

The fact that relations are implemented in terms of continuous weights does however point to a way in which certain variables might come to be seen as more important than others. This would occur when some of the weights are updated in such a way that the contribution of some of the input variables ends up being greater than others'. Such a process of weight updating might in principle occur "naturally", as a subject refines their knowledge by contact with experience. For the class of causal selection judgments that is our focus here, however, we assume (in agreement with existing counterfactual theories of causal selection) that the relevant "experience" in that regard occurs *internally*, in the form of *counterfactual simulations* by which people consider alternative scenarios for what could have happened.

In this section, we formalize counterfactual simulation as a Markov Chain Monte Carlo (MCMC) sampling process over the network's state space. In this process, agents iteratively propose small deviations

from the present state of the network (flipping one input node's activation) and accept/reject these changes probabilistically based on:

1. The **normality** of events, which is captured directly in the form of biases put on the input nodes.

2. A notion of **structural coherence**, which penalizes changes that would lead to a change in hidden layer activations.

This latter constraint will account for an independently observed feature of counterfactual simulations, their *anchoring to the real world*, as a byproduct of the inherent autocorrelation in the sampling process. In the same breath it will also capture additional effects of logical structure on the sampling propensities associated with different scenarios.

4.4.2 Desiderata: the factors driving the sampling process

Counterfactual theories hold that, when they elaborate a causal explanation for some occurrence of events, people start by mentally simulating alternative scenarios to explore how outcomes might change under different circumstances. Importantly, people are not neutral with respect to the scenarios that they consider, but they exhibit consistent preferences. Three patterns of preferences in particular will interest us here:

1. A preference for normality: people are more likely to consider alternatives that they see as more normal.

2. A preference for proximity to the real-world: people are more likely to consider scenarios similar to what actually happened.

3. Additionally, we want to consider how the preference is structured by certain *clusterings of variables*, in that certain groups of events will tend to be varied together as people consider alternatives.

Consider the following fictional example. Alice invited many people to a party. There are two groups of people that she hopes will be there in particular. One group, 'the Boys' is comprised of John, Bill, and David. Another group, 'the Girls' comprises Mary, Sue, and Claire. Alice cares about these individuals in particular because she knows that the ambiance at the party crucially depends on them. Specifically, she knows that the party will be fun if all of the boys or all of the girls are present together. Her knowledge about their contribution is captured by the following logic program:

$$LP_{party} = \{F \leftarrow J,B,D;\ F \leftarrow M,S,C\}$$

where the atoms J,B,D, etc. respectively track the occurrence of events "John comes to the party", "Bill comes to the party", and so on. This program can be CILP-translated into a neural network, yielding the network represented in Figure 4.5.

While fun at the party depends on the joint presence of boys or girls, Alice knows that each guest's probability of coming is independent as they do not coordinate. She expected each of them individually to be very likely to come, with the exception of John, whom she believed was only 50% likely to come. Suppose now that *all* six of her guests turned up in the end, leading to a fun party as expected.

Given these facts we want to characterize the alternative scenarios that Alice is likely to consider with respect to the question of which guests join the party. To help characterize those with precision, let a scenario S_i be a valuation over the binary events $\{J,B,D,M,S,C\}$, a mapping from each of these atoms to a value in $\{-1,1\}$ such that e.g., $J_i = 1$ indicates that John attends the party in S_i, while $J_i = -1$ indicates that John

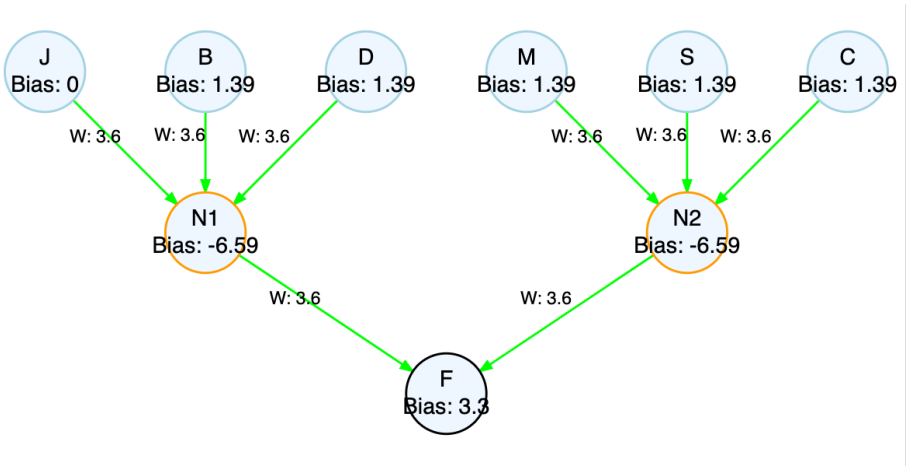


Figure 4.5: Network N_{party}. The hidden layer structure mirrors the clause structure of the logic program LP_{party}, in that each of the hidden neurons $N1, N2$ capture whether respectively "the boys" or "the girls" are present in full. The biases on the input values reflect the normality associated with the individual events. For simplicity we may assume that normality in this scenario merely reflects the degree to which each guest was expected to come (that is, there is no moral standard that is broken or enforced by each event). A bias of 0 on J reflects the prior expectation that John was 50% likely to come. A bias of 1.39 reflects an expectation that each is 80% likely to come. Biases are related to normality values via the sigmoid function, for reasons we detail below.

does not attends the party. Let $\Omega = \{-1,1\}^6$ is the set of all possible such valuations over $\{J,B,D,M,S,C\}$ and let $S_0 \in \Omega$ be the present context of reference (what actually happened), that is the scenario in which $J = B = D = M = S = C = 1$.

On these premises, the relative accessibility of scenarios can be denoted semi-informally as a sampling propensity $SP(\Omega, S_0)$ (in the sense of Icard 2015), that captures the extent to which each scenario $S \in \Omega$ is prone to come to Alice's mind as she simulates alternatives from a reference scenario S_0. The higher $SP(S_i, S_0)$, the more readily will the scenario S_i come to Alice's mind as she considered alternatives to the present-world scenario S_0. This enables us to spell out the three patterns of preferences pointed in terms of order relations between scenarios that we expect the measure SP to follow, as I do below. I'll use the notation $S_0[\neg J]$ as a shortcut to denote the scenario in which every event that occurred in S_0 occurs, except for the fact that John is now absent (and similarly with other scenarios and event variables).

Normality preference. Let's assume for simplicity that moral/conventional notions of normality don't play a role here, so that the normality of each event can be reduced to Alice's prior expectation about those events. Given that every guest was equally expected to come, except for John, who was less excepted, the scenario in which John doesn't show up comes to mind more readily than the scenario in which Bill doesn't show up. In other words we have:

$$SP(S_0[\neg J], S_0) > SP(S_0[\neg B], S_0)$$

Similarity to the real-world. The preference for alternatives closer to the real-world means that, everything else being equal, events that have occurred in the current scenario of reference will be more likely to occur in alternatives that we consider for it, than events that did not occur. For example, an alternative in which two of the boys don't show up is less prone to come to mind than an alternative in which only one of these boys don't show up. This gives us:

$$SP(S_0[\neg J], S_0) > SP(S_0[\neg J][\neg B], S_0) \quad \text{and} \quad SP(S_0[\neg B], S_0) > SP(S_0[\neg J][\neg B], S_0)$$

Clustering of variables. On top of that, it would seem (intuitively), that the way in which we explore the space of alternatives is itself organized by the way in which the relevance of event is structured by one's causal knowledge. In our example, this would be reflected in the intuition that, for instance, the alternative scenario in which all three boys are absent is easier to access that the one in which two of the boys and one of the girls are absent. It seems intuitively easier to "remove" in imagination all of the boys from the party in one feel swoop, whereas removing two boys and one girl seems like more work. This gives us:

$$SP(S_0[\neg J][\neg B][\neg D], S_0) > SP(S_0[\neg J][\neg B][\neg M], S_0)$$

This pattern is not captured by the notion of similarity to the real-world described above (at least not if we understand it in terms of the number of occurrences in common between the scenario of reference and some alternative scenario), since the scenarios $S_0[\neg J][\neg B][\neg D]$ and $S_0[\neg J][\neg B][\neg M]$ differ from S_0 each by the value of three event variables (each of which are associated with the same normality values). It is not clear that it has to do with the fact that the value of the outcome is changed in $S_0[\neg J][\neg B][\neg M]$ either. We could imagine a slightly different story where another invitee Ann is present, who is not part of either of the two groups but who can bring fun to the party on her own, and the introspective contrast between $S_0[\neg J][\neg B][\neg D]$ and $S_0[\neg J][\neg B][\neg M]$ would remain.

Hence I will assume that this contrast is driven in substantial part by the fact that the clusters $\{J, B, D\}$ and $\{M, S, C\}$ each connect to the outcome via different clauses of LP_{party}, or, equivalently, different paths in N_{party}.

4.4.3 A sampling procedure over neural networks

Having presented the different factors that are going to shape subjects' propensities to consider various alternative scenarios, we now look how these propensities can be captured via a sampling algorithm, defined over the network structures that come out of CILP-translation. The basic idea is that people imagine different scenarios and evaluate the importance of events by performing certain operations on their internal model of the relations between events and outcomes of interest. Specifically, the operations I have in mind consist in iteratively going over the following three steps:

1. Playing with the activation values of the input nodes of their neural model, resetting them to some new values.
2. Running a forward pass over the network with the new activation values.
3. Updating the weight parameters of the model.

Step 2 depends on the weights and activation functions of the network, in ways that were described in the previous sections. We do not consider step 3 for now, which is the focus of the next section. Below we look into Step 1 and the way in which it depends on the three factors described in the previous section.

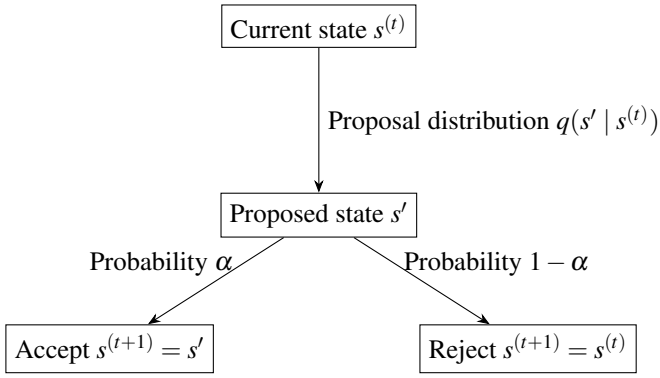


Figure 4.6: Markov Chain Monte Carlo (MCMC) state transition process. Starting from the current state $s^{(t)}$, a candidate state s' is proposed via the distribution $q(s' \mid s^{(t)})$, which (following Davis and Rehder 2020) uniformly samples states differing by one variable. The proposal is then accepted with probability α, updating the chain to $s^{(t+1)} = s'$, or rejected with probability $1-\alpha$, retaining $s^{(t+1)} = s^{(t)}$.

My proposal here is that the process by which people successively consider different scenarios can be understood in the form of a Monte Carlo Markov Chain (MCMC) sampling process, by which an agent:

1. Starts at an **initial state** $s^{(0)}$, a vector of activations over the input nodes of the network, in which the activation of each node reflects the occurrence ($= 1$) or non-occurrence ($= -1$) of events of interest in the real world.

2. Then for a number n of sampling steps, successively explore alternative states $\{s^{(0)}, \ldots, s^{(t)}, \ldots, s^{(n)}\}$, where the transition from each state $s^{(t)}$ to the next state $s^{(t+1)}$ is determined by the following two sub-steps (depicted in the decision tree in fig. 4.6):

   (a) **A proposal step**, which can be understood as sampling an alternative state s' from a proposal distribution $q(s' \mid s)$. Taking inspiration from Davis and Rehder (2020)'s *Mutation sampler* model, I assume that the proposal distribution is a uniform distribution over all possible alternative states s' that differ from the present state by the value of one variable.

   (b) An **acceptance step**, where the proposed state is accepted with an acceptance probability α (more on which below), in which case the network moves to the state $s^{(t+1)} = s'$, or rejected (with a probability $1-\alpha$), in which case $s^{(t+1)} = s^{(t)}$.

Setting the proposal distribution to be the uniform distribution over states that differ from the current one by the value of one variable only implements an idea that reasoners progress cautiously in their exploration of alternatives. To go from the original scenario of reference to a very exotic alternative, they have to move through a series of intermediate states each of which is in a relation of direct proximity to the next. The incremental sampling of scenarios closely aligns with Lewis's (1973) notion of evaluating counterfactuals in the nearest possible worlds to the actual scenario, providing a cognitively plausible anchor in real-world situations. The anchoring of simulations to the current scenario of reference will be cemented by the way in which we set the acceptance probability α. I will propose a way to determine the acceptance probability α for any proposed state s' to be accepted as the new state from any state $s^{(t)}$ in such a way as to confer the following high-level trends to the sampling process:

- It will be sensitive to the **normality** of the individual events. The probability that each proposal be accepted depends on a bias θ attached to each input variable. This bias is meant to represent the normality that a subject attaches to a certain event in the form of a continuous variable. In the context of experiments where prescriptive influences on normality are controlled, we may assume that normality reduces to the probability $P(E)$ explicitly associated with certain vignettes (e.g. in the form of the proportion of colored balls in an urn, as in the experiment presented in the previous chapter). In such cases one may want to determine biases as a function of probabilities, via the logit function

$$\log\left(\frac{P(E)}{1-P(E)}\right). \tag{4.16}$$

  This way of setting biases ensures that the long-term sampling frequencies over input variables converges to the marginal probabilities $P(E)$, given additional assumptions that I spell out below. For this reason it provides a convenient (a common) default assumption — which I endorse for the sake of modeling the experimental results from the experiment in chapter 3 in the supplementary materials attached to this dissertation. This assumption remains however arbitrary, inasmuch as there is no direct proof that subjects internalize the probabilities provided *exactly*.

- It will exhibit a strong autocorrelation, meaning that successive states will look alike, more so than one would expect from independent samples. This is a common feature of many MCMC sampling processes. The implementation I propose below makes the autocorrelation particularly strong, by assuming that the acceptance probability is influenced by the degree to which the new activation value proposed for an input node aligns with the activation values of the hidden nodes to which that node is connected in the current state $s^{(t)}$. This assumption generates auto-correlation by discouraging transitions to states that will lead to a change in the hidden nodes' values after a forward pass. For example, in the example presented in the previous subsection, going from the input state corresponding to S_0 (the scenario in which all guests are present) to $S_0[\neg J]$, the scenario in which all but John are present is discouraged because this would change the value of the hidden node $N1$ after forward pass.

  This auto-correlation, whenever the number of sampling steps is small, would lead the samples taken to resemble the initial state more so than they would if each sampling step was independent from the previous.

- Importantly, the autocorrelation dynamic generated will not affect all changes uniformly but will "hinge" around clusters of variables corresponding to the common causes of each hidden node of the network. Indeed, if the discouraged move from S_0 to $S_0[\neg J]$ ends up getting accepted (in spite of its low acceptance probability), subsequent moves that involve flipping the value of other nodes in the "boys" cluster from 1 to -1 are no longer discouraged (in fact, they are encouraged), since now these moves align with the new value of $N1$ after forward pass.

  This cashes out the intuition that the scenario in which al three boys are absent is easier to imagine that the one in which two of the boys and one of the girls are absent.

These dynamics will be generated by setting the acceptance probability α to be determined as a function of two *loss profile* terms $\Lambda_s^{(t)}$ and $\Lambda_{s'}^{(t)}$ associated respectively with the current state $s^{(t)}$ and the proposed state $s'^{(t)}$, each of which is a sum of two loss terms, λ_{input} and λ_{hidden}, described below.

An *Input Loss* term λ_{input}. This denotes the loss value associated with the current value of the input node i proposed for flipping at each step. Similarly, I will use $\lambda_{\text{input}'}$ to denote the loss value associated with the *proposed* value of the node proposed for flipping. Both loss terms entirely depend on the bias attached to the node in question. We want it to be such that the higher the bias on a node, the higher the acceptance probability of states where that node is "on" (i.e. has an activation value of 1). One possibility, which I adopt for convenience, is the negative log-likelihood function:

$$\lambda_{\text{input}} = L(x_i, \theta_i) = -\left[x_i \cdot \log(\sigma(\theta_i)) + (1 - x_i) \cdot \log(1 - \sigma(\theta_i))\right] \tag{4.17}$$

with x_i equal to a_i the current activation of the node proposed for flipping, remapped onto the interval $[0,1]$ via $x_i = \frac{a_i+1}{2}$. The equation for $\lambda_{\text{input}'}$ is the same but taking $x_i' = \frac{a_i'+1}{2}$ as argument. In both cases the θ_i argument does not change however and is handled by:

$$\sigma(\theta_i) = \frac{1}{1+e^{-\theta_i}}, \text{ the sigmoid function.} \tag{4.18}$$

One advantage of this option (which makes it a convenient default assumption) is that it makes the λ_{input} loss term exactly proportional to the marginal probability $P(E)$ of the flipped node, assuming biases have been set by eq. (4.16) above. This is easy to see, in that the logit transformation in 4.16 is canceled by the sigmoid in eq. (4.18), so that λ_{input} ends up being exactly proportional to $P(E)$. There is however no reasons to believe that people are able to internalize the probabilities of events with this kind of precision. Available evidence on the impact of event probabilities on judgments (e.g. Morris et al. 2019) merely indicate that subjects' judgments are sensitive to incremental changes in suggested event frequencies (e.g. the presence of one more colored ball in an urn), but those could be accounted by any other loss measure that makes λ_{input} an increasing function of θ. Interestingly, using other such measures, like for example the squared error function

$$L(a_i, \theta_i) = (a_i - \tanh(\theta_i))^2 \tag{4.19}$$

would result in long-term sampling frequencies for events that *ordinally* track probabilities $P(E)$, yet do not match them. This (or other loss measures) could provide an interesting modeling base to implement non-standard measures of uncertainty, such as the imprecise and comparative measures presented by Ding, Holliday, and Icard (2021) and Holliday and Icard (2013), although we don't explore that potential here.

A *Hidden Loss* term λ_{hidden}. This loss term tracks the distance between the activation of each hidden node under the current state and its value as binarized via the threshold A_{min}. It is computed via the negative log-likelihood function as we do for λ_{input} in eq. (4.17) (although similar remarks on this choice as those made above for λ_{input} apply here). Similarly, $\lambda_{\text{hidden}'}$ tracks that distance for the proposed state values. That distance will always be very small for current state values (meaning that λ_{hidden} will also be very small), since the current value of hidden nodes has been determined by the current value of the input node at the last forward pass. Take for example in the network N_{party} presented as example above in the initial state where all inputs have activation value 1, and suppose we propose to flip the activation node J to -1. The total input to $N1$ in the current state is 4.2, which corresponds to a loss of $-\log(\sigma(4.21)) \approx 0.015$.

By contrast, the distance will be very large every time the proposed flip would lead to a change in the hidden value to which the target input neuron is connected. To use the same example, in the proposed state where $J = -1$, the input to $N1$ is -3, which gives a loss of $-\log(\sigma(-3)) \approx 3.05$.

Note that we only need to compute the loss over the hidden variables to which the node proposed for

flipping is connected. This is because (for reasons explained below) we later subtract the λ_{hidden} and $\lambda_{\text{hidden}'}$ terms from one another so that whatever the losses over other hidden nodes are they will cancel out in that subtraction. This is a factor of efficiency for the algorithm proposed here, which is a feature of the one-flip-at-a-time principle that motivated its initial adoption in the Mutation Sampler model (Davis and Rehder 2020). Both λ_{hidden} terms are then also divided twice:

1. By the global weight parameter W of the model coming out of translation. Remember from the last section that the weight parameter W is determined as a function of the threshold A_{min}, and that the CILP algorithm allows us a margin of freedom to decide what the value of A_{min} is going to be. The higher we set A_{min}, the bigger the absolute value of W and the closer the activation dynamics of the network get to the "all-or-nothing" truth-functional behavior of the source program. Yet we don't want the level of autocorrelation in the sampling to depend on the (partly arbitrary) choices we make with respect to A_{min}. Dividing λ_{hidden} terms by W makes sure that whatever choices we make there, their influence on λ_{hidden} will be absorbed by that division.

2. By the number of hidden neurons to which the node proposed for flipping is connected (in the present case there is only one). This prevents the process from being too strongly autocorrelated (to the point of stasis) in contexts involving more complex rules where each input is connected to a large number of hidden nodes.

These divisions would turn the loss values of ≈ 0.015 and ≈ 3.05 computed above for respectively the initial state $s^{(0)}$ and proposed state $s'^{(0)}$ into loss terms

$$0.015/3.6 \approx 0.004, \text{ and } 3.05/3.6 \approx 0.85.$$

Summing and subtracting loss terms. From here, the acceptance probability of the proposed state can be computed by going over the following steps.

1. Sum loss terms to obtain global loss profiles Λ_s and Λ'_s for the current and proposed states:

$$\Lambda_s = \lambda_{\text{input}} + \lambda_{\text{hidden}} \tag{4.20}$$

$$\Lambda_{s'} = \lambda_{\text{input}'} + \lambda_{\text{hidden}'} \tag{4.21}$$

2. Subtract Λ_s and $\Lambda_{s'}$ to obtain Δ_{proposal}, a term that tracks how attractive the new proposed state is relative to the current one:

$$\Delta_{\text{proposal}} = \Lambda_s - \Lambda_{s'} \tag{4.22}$$

   In general, Δ_{proposal} will be larger when the proposed flip moves a node in a direction opposite to its normality bias, and when it would lead to changes to the values of hidden nodes.

   In the example above, where the proposal is about flipping J, which has a bias of 0, λ_{input} will be the same in both the current and proposed state (i.e. $\lambda_{\text{input}} = \lambda_{\text{input}'}$, hence Δ_{proposal} will reduce to:

$$\Delta_{\text{proposal}} = \lambda_{\text{hidden}} - \lambda_{\text{hidden}'} = 0.004 - 0.85 = -0.846$$

   The loss difference Δ_{proposal} can then be converted into an acceptance probability α by taking:

$$\alpha = \min\left(1, \exp(\Delta_{\text{proposal}})\right) \tag{4.23}$$

The exponential term ensures Δ_{proposal} is a positive real number. That number will systematically be greater than 1 every time Δ_{proposal} is positive, meaning that states more attractive than the current one will systematically be accepted as soon as they are proposed. Negative Δ_{proposal} values, on the other hand, will yield a probability between 0 and 1, which gets lower as Δ_{proposal} gets lower. In our running example, we would have

$$\exp(-0.846) \approx 0.43$$

meaning that the proposal of $S_0[\neg J]$ from the initial would be less than half likely to be accepted.

4.4.4 Postcript: on probabilities, sampling propensities, and mere propensities

More details about the inner workings of the sampling algorithm described here are provided in the code accompanying this dissertation. Before moving on to the next section on weight updates, I make a few high-level remarks on the process.

The way in which we have set the acceptance probability above guarantees that the stationary distribution $\pi(s)$ of the procedure over states s of the network will be proportional to the term $\exp(\Lambda_s)$. Meaning that that, in the long run, the fraction of time the Markov chain spends in each state s converges to its stationary distribution:

$$\pi(s) \propto \frac{1}{Z}\exp(\Lambda_s), \tag{4.24}$$

where Z denotes the *partition function*, defined as:

$$Z = \sum_{s' \in S} \exp(\Lambda'_s). \tag{4.25}$$

In other words, $\exp(\Lambda_s)$ plays a role equivalent to an *Energy function* for the sampling process. The features of the probability distribution $\pi(s)$ that it induces can therefore be described as a function of the two terms λ_{input} and λ_{hidden} in which Λ_s decomposes:

- On the one hand, the term λ_{input} guarantees that the *marginal probability distribution* over individual node activations will be sensitive to the biases θ on these nodes.
- On the other hand, the term λ_{hidden} skews that distribution towards states in which input nodes that connect to the same hidden node will tend to have the same activation. In other words, it introduces correlations between the activation of nodes that have a joint effect on the outcome.

The skew introduced by the latter process will bias the distribution $\pi(s)$ away from the assumption of independence between input variables; meaning that it will introduce dependencies between them even when they are not given by the data. This indicates one way in which people's sampling propensities over alternative scenarios will differ from the objective probability distributions over events reflected in the data given to them.

Another difference, already mentioned above, emerges from the autocorrelation of the process. Assuming — as we will — that the number of sampling steps is small, would lead the scenarios considered to resemble the real-life scenario of reference, more so than they would if each sampling step was independent from the previous.

Finally, sampling propensities might diverge from objective probabilities if the biases on the individual nodes do not reflect those directly (this would be typically be the case if prescriptive notions of normality

are involved), or if the function that turns biases θ into loss terms λ_{input} is not one (like the log-likelihood function above) that preserves the probabilistic information.

These facts make the whole procedure a good place to highlight the relations between three notions: *objective frequencies* on the one hand, which are provided by the real-world or the experimenter; *sampling propensities*, on the other, here represented by the loss term Λ_s. The distribution that they induce may match objective frequencies, but would only do so in the special case where we fix $\lambda_{\text{hidden}} = 0$ and where biases track the frequencies of events. Finally, *propensities* themselves, understood in the sense of Popper (1959) as

> *"weights" of possibilities which are endowed with dispositions to realize themselves, and which are taken to be responsible for the statistical frequencies with which they will in fact realize themselves in long sequences of repetitions.*

and which in our case come in the form of the weights W and biases θ of the network, which are the dispositions by which neuron activations realize themselves and one another, and that give rise to the sampling propensities over alternative scenarios that we later observe.

These distinctions give us the right framing to highlight the nature of the account that I will propose in the last two sections. Existing counterfactual theories explicate the causal importance of events as a function of the *sampling propensities* of scenarios in which they are the crucial cause of the outcome — whether they explicitly frame it in these terms as Icard, Kominsky, and Knobe 2017 or not. The account I want to propose next instead proposes to understand causal importance in terms of the *propensities themselves* by which some events bring about others in one's internal model.

4.5 An update mechanism on the samples generated: Layer-Wise Feedback Propagation.

4.5.1 Intermediate summary and presentation of this section

In the preceding sections, I have introduced the framework of logic programs, as a tool for highlighting elements of the structure of our internal programs for representing causal relations that SCM representations tend tosa overlook.. I have shown how these programs can be faithfully translated into neural networks via the CILP translation algorithm. These neural networks constitute internal generators that can be used as *sampling instruments* for counterfactual scenarios that respect known causal relations between events. The sampling algorithm presented in the previous section achieves this by exploring alternative input configurations, prioritizing scenarios that align with event normality as well as the logical groupings implied by the hidden layer connectivity of the network.

While these sections explain *which* counterfactuals are considered, they do not yet account for *how* these simulations translate into graded causal judgments—why some causes feel more explanatory than others, even when equally necessary or sufficient to the outcome. The initial CILP translation assigns uniform weights to all connections, yet human causal judgments exhibit systematic asymmetries in their preferences for events that are equally connected to the outcome in the relevant model.

In this section, I present a weight-update procedure for these neural networks, by which such asymmetries are introduced. The procedure is based on *Layer-Wise Feedback Propagation (LFP)* (Weber et al. 2025) — a neurally plausible credit assignment scheme that propagates rewards backward through feedforward networks. Unlike gradient backpropagation (Rumelhart, Hinton, and Williams 1986), which is about minimizing the distance between a network's prediction and some assigned target, LFP is about allocating credit to each parameter of the network as a function of its relative contribution to a certain outcome.

My aim in the pages below is to present a version of LFP that captures explanatory behaviors involved in causal selection. First, in **section 4.5.2**, I present LFP in the original context of its development, as a scheme for training neural networks for tasks of prediction or classification. Although originally developed as a neural network training method, LFP naturally lends itself to modeling cognitive processes involved in causal explanations, as it explicitly captures the notion of relative contributions of network components to an outcome. In **section 4.5.3**, I show how that same scheme can be repurposed as a mechanism for emphasizing the contribution of the variables most relevant to the outcome in each of the alternative scenarios explored by the counterfactual simulation process presented in the last section. Finally in **section 4.5.4** I show how such a mechanism would result in high-level dynamics of "abnormal inflation" at the level of input-to-hidden connections (whereby the weights on connections coming from abnormal input events are rewarded more so than those coming from normal ones) and "abnormal deflation" at the level of hidden-to-output connections, which can account for the corresponding patterns of causal selection judgments. The next section (**Section 4.6**) will then show the updated weights obtained in this way effectively translate in causal importance scores for the corresponding variables.

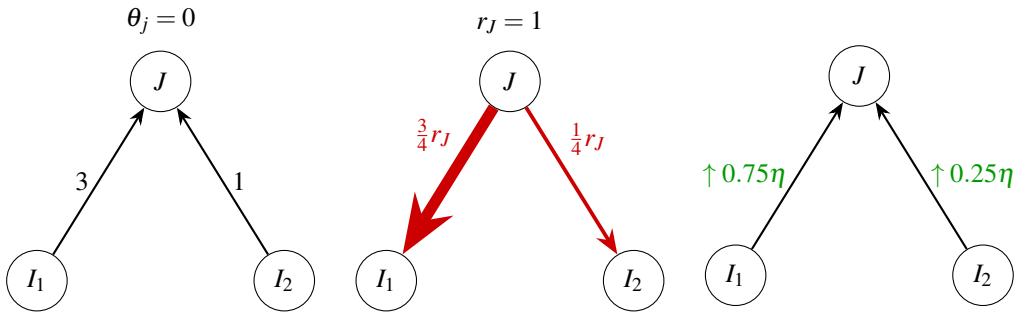
4.5.2 A weight update method for neural networks: Layer-wise Feedback Propagation

Layer-wise Feedback Propagation (LFP) denotes a class of methods recently developed (Weber et al. 2025) for training neural networks. It takes direct inspiration from the better known Layer-Wise **Relevance** Propagation (LRP) methods (Bach et al. 2015; Montavon et al. 2017; Montavon, Samek, and Müller 2018), whose original purpose was to explain the predictions of deep neural networks by decomposing the contribution to the output through successive layers all the way back to input features. I do not discuss LRP methods in detail here, as I will do so in section 4.6, where those play a role in accounting for how network weights translate into causal importance scores. Instead, I present LFP methods directly. In this subsection, I first briefly present them in the context of their original use (training neural networks for prediction/classification).

LFP as training scheme. LFP is a training method for neural networks based on a backpropagation procedure. It is similar to gradient backpropagation (Rumelhart, Hinton, and Williams 1986) in that it runs a backward pass over the network to assign credit to parameters. Starting with the output layer of the network, it first assigns credit to neurons and parameters in the layer $l-1$ that immediately precedes it. The credit values assigned to $l-1$ then serve to assign credit to neurons and parameters in $l-2$ and so on until the input layer of the network is reached.

The main difference with gradient propagation consists in the way in which credit is assigned. The simplest way to characterize the difference is the following. Gradient backpropagation is about *loss gradients* and *post-activation values*. This means that it will update parameters at layer $l-1$ in the direction that minimizes the distance between the activation of neurons at layer l and some objective. On the other hand, LFP is about *relative contributions* and *pre-activation inputs*. This means that it will reward parameters and neurons at $l-1$ as a function of the relative share of the net input to neurons in l that each contributes. To do so, it will normalize the contribution of one neuron to another by the sum total of contributions to the target neuron. To make this clear, consider a feedforward neural network with activations a_i and weights w_{ij} connecting neuron i at layer l to neuron j at layer $l+1$. The *net input* to a neuron j in this network is given as:

$$z_j = \sum_i w_{ij} a_i + \theta_j, \tag{4.26}$$



(a) Forward pass: $z_J = 3a_1 + 1a_2$
Inputs $a_1 = a_2 = 1$

(b) Reward distribution: $r_i = \frac{z_{iJ}}{z_J} r_J$
$r_1 = 0.75$, $r_2 = 0.25$

(c) Weight updates: $\delta w_{iJ} = \frac{|w_{iJ}|a_i}{z_J} r_J$
Proportional to rewards

Figure 4.7: Layer-wise Feedback Propagation mechanics. (a) Initial network with input nodes I_1, I_2, contributing differently to neuron J. (b) Reward propagation backward following Eq. 4.28, proportionally allocating credit based on relative contributions. (c) Weight updates via Eq. 4.30: updates scale with reward magnitude.

where θ_j is the bias term for neuron j. Activation values follow a convention (inherited from CILP translation) such that $+1$ indicates an event's occurrence and -1 its non-occurrence. Positive weights represent supporting evidence, negative weights inhibitory evidence. The *contribution* z_{ij} of neuron i to the net input of neuron j is then given as:

$$z_{ij} = w_{ij} a_i. \tag{4.27}$$

And the *relative contribution* of neuron i to the net input of neuron j will correspond to the fraction

$$\frac{z_{ij}}{z_j}.$$

This notion of relative contribution serves as a basis for the simplest LFP propagation technique. Let's assume that each neuron in the output layer of the network has been assigned a reward score r_o by some function (which we'll define later). One can then propagate rewards from neurons j at every layer of the network to neurons i at layer $l-1$ in a way directly proportional to their relative contribution by:

$$r_{ij} = \frac{z_{ij}}{z_j} \cdot r_j, \tag{4.28}$$

where r_{ij} is the reward that neuron i gets from neuron j.

The process is represented in figs. 4.7a to 4.7b. Suppose that neurons I_1, I_2 both have activation values $a_{I_1} = a_{I_2} = 1$ and connect to J with positive weights of $w_{I_1 \to J} = 3$ and $w_{I_2 \to J} = 1$. There is no bias on J ($\theta_j = 0$) and the reward r_J is given as $r_J = 1$. Here I_1 is contributing 3/4 of the net input to J while I_2 is only contributing 1/4, hence following the logic of eq. (4.28), the reward r_J will be "spread" in a 3-to-1 proportion between I_1 and I_2.

As rewards are propagated in this way, rewards r_i associated with neurons at every layer after the output layer are assigned by summing over the reward fractions inherited from neurons at layer $l+1$:

$$r_i = \sum_{j \text{ at } l+1} r_{ij}. \tag{4.29}$$

This means that, for example, if neuron I_1 of Figure 4.7 were to also inherit a reward value of 0.5 from some other neuron J_2 in the same layer as J, it would gather a total reward $r_{I_1} = 0.75 + 0.5$. These new reward scores are then used to prolong the propagation process to layers further down, until all the neurons of the network have been assigned a reward score.

After, or at the same time as reward scores are assigned to neurons, weights are updated in proportion to the amount of reward they channel to neurons at previous layers. The simplest way to do so is via the equations:

$$\delta_{w_{ij}}^{\text{lfp}} = \frac{|w_{ij}| \cdot a_i}{z_j} \cdot r_j \tag{4.30}$$

$$w_{ij}^{\text{new}} = w_{ij}^{\text{old}} + \eta \cdot \delta_{w_{ij}}^{\text{lfp}} \tag{4.31}$$

where η is a learning rate parameter. Equation 4.30 essentially amounts to replacing w_{ij} in z_{ij} from Equation (4.28) with $\text{sign}(w_{ij}) \cdot w_{ij} = |w_{ij}|$. The weight w_{ij} is multiplied with the parameter sign, following the convention:

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ -1 & \text{otherwise.} \end{cases} \tag{4.32}$$

This means that every edge outgoing from a neuron will be rewarded (in terms of updates) in direct measure to the size of the reward that it "brought home" to that neuron in the backward reward propagation process described by eq. (4.28), as illustrated in fig. 4.7c. The play of signs ensures that the update will go in the direction that would bring more of that reward at the next iteration.

Biases θ_j on the individual neurons can also be updated using Equation 4.30. This would basically amount to treating biases as if they were weight inputs coming from a neuron b with a constant activation value $a_b = 1$ and a weight w_{bj} equal to θ_j.

Handling negative activations. Equations (4.28) to (4.31) above constitute the simplest LFP scheme possible, which makes it useful for the purpose of illustrating the general principles behind the procedure. It is in general well-behaved for networks that only involve positive activations. When negative activations are involved, however, an issue of exploding feedbacks can arise if $|z_{ij}| \gg |z_j|$. Suppose for example that the weight $w_{I_2 \to J} = 1$ in fig. 4.7a is replaced by $w'_{I_2 \to J} = -2.99$. This makes the net input z_J equal to $w_{I_1 \to J} - w'_{I_2 \to J} = 0.01$. As a result, the rewards propagated to I_1 and I_2 by eq. (4.28) rise to 300 and 299 respectively. Large values like these render training highly unstable. To alleviate this problem, variants of LFP can be proposed that separate positive and negative contributions to a given neuron. One such variant is known as $\text{LFP}_{\alpha\beta}$:

$$r_{ij}^{\alpha\beta} = \begin{cases} \alpha \dfrac{z_{ij}}{z_j^+} \cdot r_j & \text{if } z_{ij} \geq 0, \\ -\beta \dfrac{z_{ij}}{z_j^-} \cdot r_j & \text{else.} \end{cases} \tag{4.33}$$

with:

$$z_j^+ = \sum_i \max(z_{ij}, 0), \quad z_j^- = \sum_i \min(z_{ij}, 0), \quad \text{and } \alpha - \beta = 1$$

Separating positive and negative contributions in Equation (**??**) prevents unstable feedback loops arising from large opposing contributions. It thereby ensures stable propagation of explanatory signals throughout the network. The $\alpha - \beta = 1$ constraint ensures that the sum of rewards over neurons of a given layer is

conserved across layers. Meaning that if the output layer is assigned a reward of 1, then the sum of rewards accumulated on neurons of every layer after the backward pass will also be equal to 1. The LFP$_{\alpha\beta}$ rule is directly inspired from the exactly analogous LRP$_{\alpha\beta}$ rule for Relevance propagation. It makes sense for rewards to be conserved across layers in the context of Relevance propagation. It captures the intuition that the relevance accumulated over input neurons of the input layer tracks exactly their share of responsibility in the final outcome. For a weight-update mechanism like LFP, however, this constraint is not as indispensable and it would make sense for some layers to incur stronger updates than others for certain observations. Weber et al. (2025) propose to relax it in consequence, with the following alternative scheme, which separates positive and negative contribution but does not enforce such constraint, which they dub LFP$_{z^+z^-}$:

$$r_{ij}^{z^+z^-} = \begin{cases} \dfrac{|z^+|}{|z_j^+|+|z_j^-|} \cdot \dfrac{z_{ij}}{z_j^+} \cdot \text{sign}(z_j) \cdot r_j & \text{if } z_{ij} \geq 0, \\ -\dfrac{|z^-|}{|z_j^+|+|z_j^-|} \cdot \dfrac{z_{ij}}{z_j^-} \cdot \text{sign}(z_j) \cdot r_j & \text{else.} \end{cases} \tag{4.34}$$

$$\delta_{w_{ij}}^{z^+z^-} = \frac{|w_{ij}| \cdot a_i}{|z_j^-|+|z_j^+|} \cdot \text{sign}(z_j) \cdot r_j, \tag{4.35}$$

In the toy example presented above, where $w_{I_1 \to J} = 3$ and $w'_{I_2 \to J} = -2.99$, this would mean that both I_1 and I_2 inherit a reward with about the same absolute value of $\approx r_J/2 = 0.5$, as they both represent about half of the sum of absolute input values received by J. I_1 will receive 0.5, while I_2 will receive -0.5, by virtue of the negative sign for cases where $z_{ij} < 0$. The update scheme instantiated by eq. (4.35) would result in $w_{I_1 \to J}$ and $w'_{I_2 \to J}$ both incurring an update of $+0.5\eta$, which would increase the absolute value of the positive $w_{I_1 \to J}$ but decrease that of the negative $w'_{I_2 \to J}$.

4.5.3 Layer-Wise Feedback Propagation for explanations

Weber et al. 2025 show that, LFP procedures like the above can achieve performance similar to gradient backpropagation on benchmark datasets. My interest in LFP here is different, however. It lies in the fact that it constitutes a great method for emphasizing which components of a causal system are the main *driving factors* behind a certain prediction. My account assumes that when people seek explanations for an event, they proceed as follows. First, they consider the present occurrences of events, as well as a few similar scenarios via the sampling procedure described in the previous section of this chapter. Second, in each of these contexts, they use their internal simulation model to generate predictions for the alternative causal setting that they consider in imagination. Third, they *backtrack* from the predictions thus generated to the processes that generated them, all the way to the initial causes, to see which pieces and parameters of their internal model are driving that prediction, and *tune-up* (or down) the parameters as a function of their contribution.

This backtracking process essentially consists in asking successive "why?" questions about the mechanism that leads to a certain prediction. For example, in the context of the experiment described in Chapter 3, where a player needs two purple balls or two yellow balls to win a round of the game, subjects would ask, for each scenario that they contemplate: "which color(s) led me to win (or lose)?" Then, ask about each color: "which particular draws led me to complete (or not) that color?"

To take a different example for clarity, suppose that this afternoon I caught a few fish, and also bought some groceries at the store, either of which would have been sufficient to cook a dinner. As I attempt to reconstruct the factors that enabled me to dine that night, I may first consider which of these two food sources

I should credit—if I only fish occasionally, my daily visit to the store will probably be the most salient explanation. Then, I may further ask which factors made it possible for me to catch fish and buy groceries respectively, as a means to refine my explanation. In that respect, it seems clear that whichever factors I choose to credit for e.g. my fishing performance (childhood lessons from an uncle, the passing cars that scared the fish, etc.) should not change as a function of the outcome (here cooking dinner) that they matter to. If, for example, my fishing performance also mattered to a contest I made that same afternoon with a friend, which had a different outcome (I lost the contest), this should not lead me to evaluate differently the factors that I credit for the number of fish I caught.

By this description I mean to emphasize that the weight-update process in which we engage is *compartmentalized* in ways that respect the modular nature of the relevant generative process, and that are not necessarily given by default in LFP rules.Compartmentalizing reward propagation aligns with cognitive intuitions that separate independent causal pathways. Humans intuitively assign causal importance within isolated subsets of explanatory structures rather than globally across all causal factors. For all the reasons discussed in Section 4.2–4.3, I take it that the relevant boundaries for that compartmentalization here are drawn by the hidden layer structure of the networks. This motivates an amendment to the rule in eq. (4.34) to ensure that reward values do not switch signs when going from output to hidden layer, as explained below in this subsection.

Another important fact to take into account is that, as we engage in counterfactual simulations and simultaneous weight updates, we don't do so with the aim of explaining what *could have happened* (in all of the imagined worlds) but of explaining what actually happened in the world of reference, which the explanation is about. One way to interpret the role of counterfactual simulations and weight updates in explanations is as a form of "stress-testing". We play with the settings of our internal program (here in the form of input causes) to better reveal its inner workings, then emphasize the key components via weight updates. The notion that subjects aim to explain what happened in the current world rather than what could have happened will have a direct translation in terms of the LFP processes by which we will model the changes in parameters. Simply put, it entails that whenever the outcome obtained in one alternative scenario is the *opposite* of that obtained in the world of reference, the reward assigned to that outcome (and then propagated over the network) should be negative. Below I define a reward function that ensures this is the case.

A reward function for explanations. I'll define the reward function as follows. Provided:

- $\mathsf{N} = \langle \mathsf{V}, \mathsf{E}, \mathbf{W}, \boldsymbol{\Theta} \rangle$, a neural network resulting from CILP translation, and an actual world of reference *AW*, such that:
  - V is the set of neurons, partitioned into:
    - Input neurons: V_{in}
    - Hidden neurons: $\mathsf{V}_{\text{hidden}}$
    - Output neurons: V_{out}
  - $\mathsf{E} \subseteq \mathsf{V} \times \mathsf{V}$ is the set of directed edges (connections) between neurons.
  - $\mathbf{W} = \{w_{ij} \mid (i,j) \in \mathsf{E}\}$ are the weights on the edges.
  - $\boldsymbol{\Theta} = \{\theta_j \mid j \in \mathsf{V}\}$ are the biases for each neuron.
- AW represents an observation, that is, a pair of vectors of activation values:

  - $\mathbf{a}_{\text{in}} = \{a_i \mid i \in \mathsf{V}_{\text{in}}\}$, a vector of activation values for input neurons that track the occurrences of events in the real-world context of interest.
  - $\mathbf{y}_{\text{out}} = \{a_{Oc} \mid Oc \in \mathsf{V}_{\text{out}}\}$, a vector of activation values for output neurons that track the occurrences of outcome events in the real-world context of interest. In what follows we will only look at cases where there is just a single relevant outcome Oc, so we will talk about a_{Oc} as the activation of the corresponding one neuron and $\mathbf{y}_{Oc}$ as the activation that tracks its value in the real-world.

- $s^{(0)}$ is the initial state of N, before considering counterfactual scenarios, generated by initializing V_{in} as $\mathbf{a}_{\text{in}}$ and then running a forward pass,
- $s^{(1)}, \ldots, s^{(n)}$: The counterfactual states of the network simulated via the process described in Section 4.4, after N has been initialized at $s^{(0)}$.
- $a_{Oc}^{(0)}$: The value of O_c at $s^{(0)}$. If N makes the right prediction for AW this is equivalent to $\mathbf{y}_{Oc}$ up to the level of approximation of the network that comes with the CILP translation,
- $a_{Oc}^{(i)}$: The activation value of O_c in state $s^{(i)}$,

The reward function for Oc is defined as:

$$r_{Oc}^{(i)} = \mathbf{y}_{Oc} \cdot a_{Oc}^{(i)} \tag{4.36}$$

This ensures that, whenever $a_{Oc}^{(i)}$ has a different sign from $\mathbf{y}_{Oc}$, $r_{Oc}^{(i)}$ will be negative. It captures the idea that what people seek to explain is the outcome in the real-world, rather than the individual outcomes predicted in each counterfactual world.

A modified LFP$_{z^+z^-}$ rule. The second adjustment we need to make to repurpose LFP$_{z^+z^-}$ into a process that emphasizes the driving factors behind a prediction is to compartmentalize the reward assignment process with respect to the different causal pathways distinguished by the hidden layer structure of the network. This can be done relatively simply, by modifying eq. (4.34) to remove the negative sign for cases where $z_{ij} < 0$. This gives the following modified rule:

$$r'_{ij} = \begin{cases} \dfrac{|z^+|}{|z_j^+| + |z_j^-|} \cdot \dfrac{z_{ij}}{z_j^+} \cdot \text{sign}(z_j) \cdot r_j & \text{if } z_{ij} \geq 0, \\ \dfrac{|z^-|}{|z_j^+| + |z_j^-|} \cdot \dfrac{z_{ij}}{z_j^-} \cdot \text{sign}(z_j) \cdot r_j & \text{otherwise.} \end{cases} \tag{4.37}$$

Given the reward function in eq. (4.36), this new rule ensures that the sign of the reward that is propagated onto hidden rules is always positive whenever $\mathbf{y}_{Oc}$ is positive. This is because it has the consequence that whenever sign(z_{0c}) is positive, r_{Oc} will also be, making both equations above systematically positive for hidden neurons directly connect to Oc.

It will have different consequences for negative outcomes, as I explain in the next section. This is because explaining negative outcomes will involve additional machinery, in the form of a new network path $\neg E \leftarrow\!\sim E$ that implements the closed-world assumption with respects to E. To explain $\neg E$, people first assume that their database contains all the relevant knowledge E, draw a new path in their inner model that implements that assumption, and then back-propagate rewards from $\neg E$ to E before they can trickle-down in the rest of the network. Because the relevant machinery is also relevant to other mechanisms introduced in the next section, it will be easier to introduce it there.

4.5.4 Overview of the emergent patterns: abnormal inflation, abnormal deflation

To give insight into the implications of the weight update process defined above, I present in this subsection some general patterns that we expect when that procedure is applied to a neural network N across a set of sampled states $\{s^{(1)},\ldots,s^{(n)}\}$.

I concentrate on two simple patterns, which also correspond to the most consistent patterns identified in the literature on causal selection judgments with binary variables: *abnormal inflation* and *abnormal deflation*. I must however clarify that the kind of inflation I have in mind is not about how the estimated importance input causes varies as a function of their normality and the rule that relates them to the outcome. That is, it is not about how abnormal (or normal) *causes* are deemed more important when each of several causes are individually *necessary* (or sufficient) for an outcome to happen, but rather about the inflation of the *weights* coming from abnormal causes in input-to-hidden connections, or from normal causes in hidden-to-output connections. The two are directly related: the inflation of certain weights will result in higher causal importance for the neurons associated with these weights. But the translation from one to the other will require some additional mechanism, which I present only in the next section.

Also, the notion of normality considered here is simply about sampling frequency. I mean to say that input-to-hidden connection weights get larger as the corresponding input nodes are less frequently active, and hidden-to-output connection weights get larger as the corresponding input nodes are more frequently active. Sampling frequency does relate to normality, though, as explained in section 4.4. For input neurons it directly relates to their individual normality bias; for hidden neurons, it relates to the normality of the input neurons from which they receive input.

Finally, I must emphasize that here I am only interested in *relative* patterns of inflation, that is, how certain connection weights are rewarded *more* than others at the same layer, not how the updates on a certain weight taken in isolation grow as a function of normality. In both cases, I concentrate on the simplest cases where $\mathbf{y}_{Oc}=1$, and when all of the input variables are active in the real-world.

Abnormal Deflation for $W_{\mathbf{hid}\rightarrow\mathbf{out}}$

Focusing on the weight matrix $W_{\text{hid}\rightarrow\text{out}}$ connecting hidden nodes to the output nodes, we find that weights originating from hidden nodes with *more frequent positive activations* across the sampled states $\{s_1,\ldots,s_n\}$ tend to increase more often. This result follows from the following facts.

1. Since $\mathbf{y}_{Oc}=1$, it follows from Equation (4.36) that for every sampled state s_i:

$$r_{Oc}^{(i)}=\mathbf{y}_{Oc}\cdot a_{Oc}^{(i)}=1\cdot a_{Oc}^{(i)}=a_{Oc}^{(i)}.$$

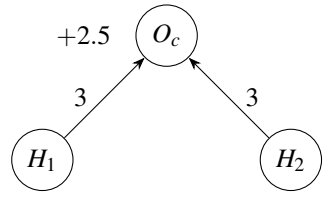
   Thus, the reward for the outcome neuron directly matches its activation in each counterfactual state.

2. The activation values a_n of neurons in N are governed by $\tanh(z_n)$. Given $\tanh(\cdot)$ only outputs activations with the same sign as its inputs, we have $\text{sign}(a_n)=\text{sign}(z_n)$. Therefore also:

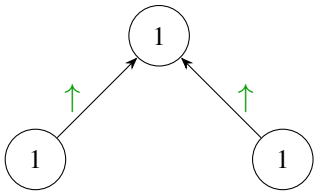
$$\text{sign}(z_{Oc})\cdot r_{Oc}=\text{sign}(z_{Oc})\cdot a_{Oc}.$$

   By substitution from the previous step, this product will always be positive.

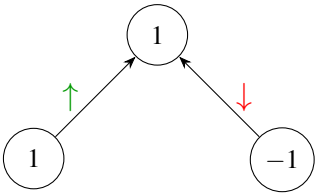
3. Given that $\text{sign}(z_{Oc})\cdot r_{Oc}$ is always positive, the sign of the weight update $\delta^{\text{lfp}}_{w_{h\rightarrow Oc}}$ (as defined by Equation (4.35)) depends solely on the sign of a_h. Whenever $a_h>0$, the update $\delta^{\text{lfp}}_{w_{h,Oc}}$ is positive, and whenever $a_h<0$, $\delta^{\text{lfp}}_{w_{h,Oc}}$ is negative.



(a) Example of a Network fragment, fresh out of translation.



(b) State $H_1 = 1, H_2 = 1, O_c = 1$: no contrasts are introduced.



(c) State $H_1 = 1, H_2 = -1, O_c = 1$: favors $W_{\text{h1}\to\text{Oc}}$.

Figure 4.8: LFP update dynamics for $W_{\text{hid}\to\text{out}}$. The subnetwork in (a) represents connections between two hidden neurons H_1, H_2 and an outcome neuron Oc. The weights and biases are such that Oc will be positively activated every time one of H_1, H_2 are, and negatively activated otherwise. (b): Whenever either is activated, its sign aligns with that of the outcome such that it will be rewarded. (c): this will introduce a contrast between weights whenever one is activated and the other isn't. In this particular case, it favors $w_{H_1\to Oc}$. Assuming that H_1 is more frequently active that H_2, this sort of situation will generate an asymmetry in weights in favor of $w_{H_1\to Oc}$ and against $w_{H_2\to Oc}$.

4. In the context of the CILP translation algorithm, the initial weights $W_{\text{hid}\to\text{out}}$ are always positive when the source is a general logic program. Thus, the magnitude $|w_{h,Oc}|$ will grow as a function of the *average positive activation* of hidden neuron h across the sampled states $\{s_1, \ldots, s_n\}$. Since activation values under $\tanh(\cdot)$ typically saturate very close to -1 or 1 in the Networks fresh out of CILP translation, the average activation effectively reflects the frequency with which h is positively active across $\{s_1, \ldots, s_n\}$.

Thus, hidden neurons that activate more frequently and selectively gain an advantage, leading to a pattern where their corresponding weights strengthen disproportionately compared to others. We refer to this effect as *abnormal deflation*, following the terminology of Icard et al. (2017). See Figure 4.8 for an illustration.

Abnormal Inflation dynamics for $W_{\text{in}\to\text{out}}$

In contrast to the pattern observed for $W_{\text{hid}\to\text{out}}$, connection weights between input and hidden nodes $W_{\text{in}\to\text{hid}}$ exhibits the opposite tendency: input nodes with the *lowest* sampling propensity across $\{s_1, \ldots, s_n\}$ end up with the strongest connections to their corresponding hidden nodes. One precision however is needed: because $W_{\text{in}\to\text{hid}}$ weights, unlike $W_{\text{hid}\to\text{out}}$, can be negative (as they are when the corresponding clause features a negative literal), here we take "lowest sampling propensity" to implicitly refer to the sampling propensity of the "desired" activation for an input node, i.e. 1 whenever the source clause features a positive literal L, and -1 whenever it is a negative literal $\sim L$. This follows from the following facts:

1. As established in the previous subsection, whenever $y_{Oc} = 1$, the reward for every hidden node h is

always positive, i.e., $r_h > 0$.

2. Since $r_h > 0$, the direction of the weight update $\delta^{\text{lfp}}_{w_{i,h}}$ (see Equation (4.35) or the relevant weight-update equation) depends solely on the sign of the product $a_{\text{in}} \times \text{sign}(z_h)$.

3. Consider the case where $\text{sign}(z_h) > 0$. Since $\text{sign}(z_h)$ indicates whether the hidden node h is active, a positive $\text{sign}(z_h)$ implies that h is actively participating in producing the observed outcome. Given the construction derived from the CILP algorithm, for h to be active, the input nodes connected to h must have activation values that are consistent with the weights $W_{\text{in},h}$. More concretely:

   - If $w_{i,h} > 0$, then $a_i > 0$ is expected to contribute positively to h.
   - If $w_{i,h} < 0$, then $a_i < 0$ is expected to align with the negative weight, effectively representing a negated literal in the corresponding logical clause.

   To avoid confusion between activation values and their logical interpretation, let us introduce a notion of *alignment*. We say that an input node i is *aligned* with respect to a hidden node h if:

   $$\text{sign}(a_i) = \text{sign}(w_{i,h}).$$

   In other words, the input node's activation matches the polarity of the connecting weight. Using this notion, $\text{sign}(z_h) > 0$ if and only if *all* input nodes connected to h are aligned with it.

4. By virtue of the relevant weight update equations (e.g., Equation (4.35)), when $\text{sign}(z_h) > 0$, all weights $w_{i,h}$ increase in magnitude by a similar amount. In such *fully aligned* cases, no differential pattern emerges between different input nodes. Thus, these scenarios do not introduce any meaningful contrast in the relative strengths of the input-to-hidden connections.

5. Now consider the case $\text{sign}(z_h) < 0$, which corresponds to the hidden node h being inactive ($a_h \approx -1$). Such inactivity arises when at least one input node is *misaligned*:

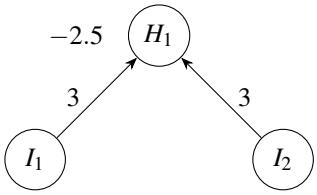
   $$\text{sign}(a_i) \neq \text{sign}(w_{i,h}).$$

   Here, two sub-cases emerge:

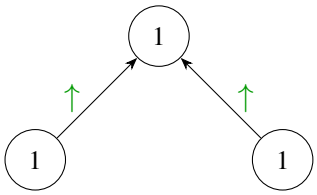
   (a) **All inputs are misaligned:** This is a symmetric inverse of the fully aligned case. All input nodes contribute negatively and are thus penalized similarly, introducing no contrasts between them.

   (b) **Some but not all inputs are misaligned:** In this mixed scenario, the weights from misaligned nodes (those responsible for driving h into inactivity) will be rewarded, as the network attempts to adjust and reduce the mismatch. Meanwhile, the weights from aligned nodes will be relatively penalized, since they failed to produce a coherent positive contribution at the hidden layer.

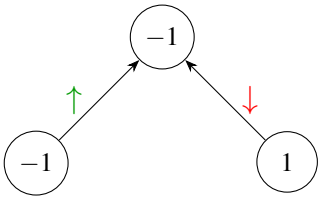
   It is in these latter, partially misaligned cases that *contrasts* in input-to-hidden weights emerge. Over multiple sampled states $\{s_1, \ldots, s_n\}$, the input nodes that are seldom aligned (i.e., those that often differ from the hidden node's desired polarity) will be those whose weights end up growing larger in magnitude.



(a) Example of a Network fragment, fresh out of translation.



(b) State $I_1 = 1, I_2 = 1, H_1 = 1$.



(c) State $I_1 = -1, I_2 = 1, H_1 = 1$.

Figure 4.9: LFP update dynamics for $W_{\text{in}\to\text{hid}}$. (a): The weights and biases are such that H_1 will be positively activated every time both of I_1, I_2 are, and negatively activated otherwise. (b): Whenever both I_1, I_2 are positively activated, their signs both align with that of the outcome so that they are both rewarded. The same is true of states where both are negatively activated. (c): this will introduce a contrast between weights whenever one is activated and the other isn't. In this particular case, it favors $w_{I_1\to H_1}$, because I_1 is negatively activated. Assuming that I_1 is less frequently active in general than I_2, this sort of situation will generate an asymmetry in weights in favor of $w_{I_1\to H_1}$ and against $w_{I_2\to H_1}$.

4.6 Defining a measure of causal importance over networks.

4.6.1 Intermediate Summary and preview of the section

The preceding sections detailed the transformation of logic programs into neural models, the exploration of counterfactual states via MCMC sampling procedures, and the subsequent weight adjustments through Layer-Wise Feedback Propagation (LFP). Collectively, these steps explain how the network's weights systematically highlight specific causes. I now want to describe how we read off a final causal importance score from the resulting, updated neural network. We have seen how "credit assignment" shapes the weights, but not yet how those updated weights translate into explicit *causal impact measures* for each candidate cause. This is what I propose to do in the present final section. I present a measure for computing causal scores that is sensitive to the structure of the networks over which it is defined. It involves two components.

The first component is a measure of **relevance** for the various input events, that tracks how important each input variable is to the outcome individually. It is computed via **Layer-wise Relevance Propagation** (LRP) procedures (Bach et al. 2015; Montavon et al. 2017; Montavon, Samek, and Müller 2019), originally developed for explaining the predictions of deep neural networks. These closely resemble the LFP procedures presented earlier, which they originally inspired. We will assume that after having updated the neural network N by which they internally represent the relevant causal relations between events into a network $\mathsf{N}^{(\text{up})}$ with new weights, subjects go back to the initial activation state (where neuron's activations track the occurrences of events in the real-world of reference). Then, they assign some Relevance value R_{Oc} onto the outcome Oc, and propagate that relevance over successive layers of the network. Relevance first flows from the outcome to hidden nodes, and then from each hidden node to the input nodes that are connected to it. The sum of relevance propagated to each layer is conserved across layers, which means that each hidden neuron in $\mathsf{V}_{\text{hidden}}$ gets assigned a certain fraction of R_{Oc}, and also each ultimately each input neuron in V_{in}, that is:

$$\sum_{i\in\mathsf{V}_{\text{in}}} R_i = R_{Oc}.$$

The amount of relevance that flows from Oc to $\mathsf{V}_{\text{hidden}}$, and from $\mathsf{V}_{\text{hidden}}$ to V_{in} is determined by the weights

on the connections through which it flows, which makes the weight asymmetries introduced by the LFP mechanisms presented in the previous section crucial.

The second component is a measure of **path complexity** that allows us to convert the relevance for a set of causal variables C into a measure of causal importance $\kappa(C,O)$ by dividing the sum of the relevance held by causes in C by a measure of the complexity of the network routes from which it gets that relevance:

$$\kappa(C,O) \;=\; \frac{\sum_{c\in C} R_c}{\mathscr{C}(C,O)}, \tag{4.38}$$

where $\sum_{c\in C} R_c$ is the sum of the relevance accumulated by the causes in C, and $\mathscr{C}(C,O)$ is a factor tracking the number of *edge-disjoint active routes* from C to O. The relevant notion of an active route here is similar to that in Hitchcock (2001a). What it does essentially is to capture the intuition that we want our explanations to be as straightforward as possible, and ideally generate all of their effect through the same mechanism. It acts like a principle of parsimony for causal selection explanations, that accounts for the fact that citing *all* of the input variables available won't deliver an optimal explanation. Only the complexity measure that is minimized through that parsimony principle is not the number of *causal entities* (or input variables) that are invoked by the explanation, but the number of mechanisms or processes that it involves. The two are presumed to be the same under the one-equation-per-variable principle that underlies SCMs, but here we have endorsed very different assumptions. The experimental data on plural cause judgments presented in Chapter 3 have moreover presented evidence that the number of causal entities is not an accurate criterion of complexity. The $\mathscr{C}(C,O)$ denominator is especially relevant for plural cause judgments (where C contains more than one variable). It will capture which groupings of those variables are naturally attractive, in ways that goes beyond the sheer impact they have on the outcome (measured by relevance), and which aren't.

The two components (relevance and path complexity measures) are conjointly relevant to the account of people's judgment presented here. I show how the measure can account for concrete patterns of judgments in sections (4.6.4)-(4.6.5). I put a special focus on **negative outcomes**, which are presented in section 4.6.5. We have seen in Chapter 3 how people's judgments for negative outcomes reveal patterns very different from those expected under the assumption that explain a negative outcome is just like explaining a positive one, but moving the target. I propose that this has everything to do with the fact that human's default representation of negation is akin to *negation as failure* rather than classical negation. Subject's internal simulation model allows them to generate instances of E via some neural machine that implements a general program like $\{E \leftarrow A,B; \quad E \leftarrow C,D\}$. But being able to generate mental simulation where E occurs does not automatically empower one to generate simulations where $\neg E$ occurs. At best, one can tell from such a program that running a simulation with certain input settings *does not bring* E about. To go from there to the notion that I have derived $\neg E$ involves the additional assumption that one's simulation model contains all the information that would have been relevant to derive E (i.e. the closed-world assumption with respect to E), which is not given by default in the general program $\{E \leftarrow A,B; \quad E \leftarrow C,D\}$.

I will introduce additional machinery that implements this assumption. It will take the form of a new path $\neg E \leftarrow\!\sim E$ which people append to their preexisting network representation. To explain $\neg E$, people add the direct path $\neg E \leftarrow\!\sim E$ to their internal model, then engage in all of the explanatory process described in this section and the previous. I show how this addition of a new path interacts with both backpropagation mechanisms presented in this section and the previous (LFP and LRP) as people backpropagate rewards and relevance from $\neg E$, as well as with the $\mathscr{C}(C,O)$ measure of path complexity.

4.6.2 Layer-Wise Relevance Propagation (LRP) for Score Computation

Layer-Wise Relevance Propagation (Bach et al. 2015; Montavon et al. 2017; Montavon, Samek, and Müller 2019) is a technique to interpret neural networks by distributing relevance scores from the output layer back to the input. It is related to LFP (for which it served as the original inspiration), but while LFP propagates *rewards* for learning, LRP propagates *relevance* for explanation. Just like with my presentation of LFP in the previous section, I start by introducing the simplest propagation rules used in the literature, because those make it most easy to convey the general idea. I then point out some intuitive limitations that justify a complexification of those rules, which I introduce next.

LRP$_\varepsilon$. One of the simplest commonly used LRP rules is the ε-rule, which propagates relevance from neurons j at layer $l+1$ to neurons i at layer l via the equation:

$$R_i = \sum_j \frac{z_{ij}}{z_j + \varepsilon} R_j \tag{4.39}$$

where R_j is the relevance of neuron j, $z_{ij} = a_i w_{ij}$ is the input contribution of neuron i to neuron j. The net input $z_j = \sum_i z_{ij}$ is the total input received by j from i and other neurons on layer l and ε is a small positive constant added for numerical stability.

Applying this rule layer-by-layer distributes some relevance score R_j at the output layer all the way down to the input layer while preserving the sum of relevance accumulated on each layer.

Application to an Example Network. To illustrate the procedure with an example, consider the neural network N depicted in Figure 4.10 and suppose that, in the world of reference AW at which relevance will be propagate, all of A, B, C all occur ($a_A = a_B = a_C = 1$).

All connection weights of the network are positive, but they differ in magnitude. The hidden-to-output weights $W_{\text{hid}\to\text{out}}$ follow the ranking $W_{H_{AB}\to E} < W_{H_{AC}\to E} < W_{H_{BC}\to E}$. At the input-to-hidden level, A tends to have stronger connections to its targets than the other inputs. These weight patterns would result in the propagation dynamics depicted in Figure 4.11. At the hidden-to-output level, relevance flows preferably to H_{BC} because it has the strongest connection to E. At the level of input-to-hidden connections, A gathers more relevant from H_{AB} or H_{AC} than respectively A or C. But these hidden nodes also happen to be the ones that hold the weakest amount of relevance. The reverse situation holds for C: it receives relevance from the hidden nodes H_{AC} or H_{BC} that hold the most, but gets a weaker share of them because of weaker input-to-hidden connections. B occupies an intermediate position, which in this case ultimately credits it with the most relevance.

Limitations with Negative Weights and Activations Just like LFP measures, the simplest propagation rules tend to encounter limiting cases. The limitation of the rule in eq. (4.39) becomes apparent when one considers contexts with negative activation values. Consider again the network described earlier, but now under a different real-world observation AW where $A = 1$, $B = 1$, and $C = -1$ (i.e., $a_A = 1$, $a_B = 1$, and $a_C = -1$). Running a forward pass on the network with these input activations yields the following hidden-layer activations:

$$a_{H_{AB}} \approx 1, \quad a_{H_{AC}} \approx -1, \quad a_{H_{BC}} \approx -1.$$

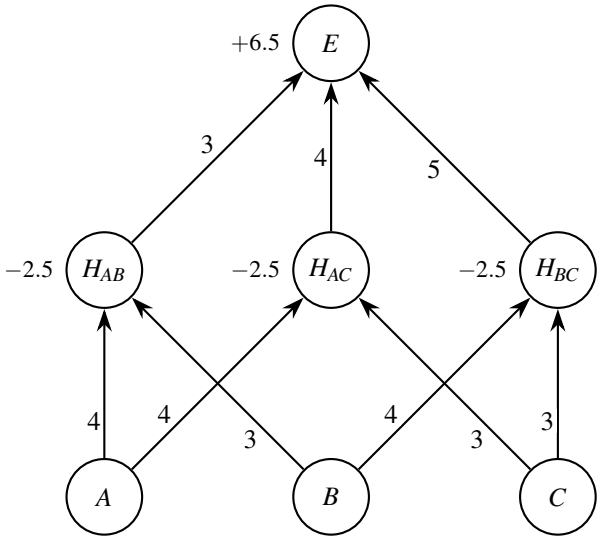


Figure 4.10: Neural network diagram with inputs A, B, C; hidden neurons H_{AB}, H_{AC}, H_{BC}; and output neuron E. The weights on the edges are indicated next to the arrows, and biases are shown next to the neurons. This network is an approximate representation of the network $\mathsf{N}^{(\mathrm{up})}$ obtained after **(i)** initializing a network N that translates the logic program $\{E \leftarrow A,B;\ E \leftarrow A,C;\ E \leftarrow B,C)$ **(ii)** Running the weight update process described in **??** with 20 samples generated from an initial state where $a_A = a_B = a_C = 1$ and with biases on the inputs reflecting the marginal probabilities $P(A) = 0.05$, $P(B) = 0.5$, $P(C) = 0.95$. These basically capture the assumptions of our the present theory for the causal setting in Quillien and Lucas (2023)'s experiment (2b), which is also the Experiment 1 of the paper presented in Chapter 3. The approximation in the weights represented on the figure involve rounding the weights (but preserve the basic proportions) and also ignore all of the updates performed on individual neuron's biases (we do update them in the weight update process, but they are ignored in relevance propagation), to enhance clarity.

We compute the total activation input to the output neuron E (without bias θ), which gives us:

$$z_E = W_{H_{AB},E} \cdot a_{H_{AB}} + W_{H_{AC},E} \cdot a_{H_{AC}} + W_{H_{BC},E} \cdot a_{H_{BC}} \tag{4.40}$$

$$= (3)(1) + (4)(-1) + (5)(-1) = 3-4-5 = -6. \tag{4.41}$$

Since z_E appears in the denominator when normalizing the relevance scores of the hidden units, the neuron H_{AB}, which is positively activated, receives a negative relevance score, while H_{AC} and H_{BC}, which are negatively activated, receive positive relevance scores. Propagating these scores all the way back to the input layer yields a counterintuitive situation in which C emerges as the most important contributor to the outcome, despite C being the only variable that *does not* actually contribute to the outcome (see Figure 4.12).

This paradoxical result arises in part because we represent events that did *not* occur using negative activations instead of activation values of zero. Since we wish to preserve this representation, we must adopt a different LRP measure. One relevant alternative (which I'll adopt) is the $\mathrm{LRP}_{\alpha\beta}$ rule:

$$R_i = \sum_j \left(\alpha \cdot \frac{(z_{ij})^+}{(z_j)^+} - \beta \cdot \frac{(z_{ij})^-}{(z_j)^-} \right) R_j \tag{4.42}$$

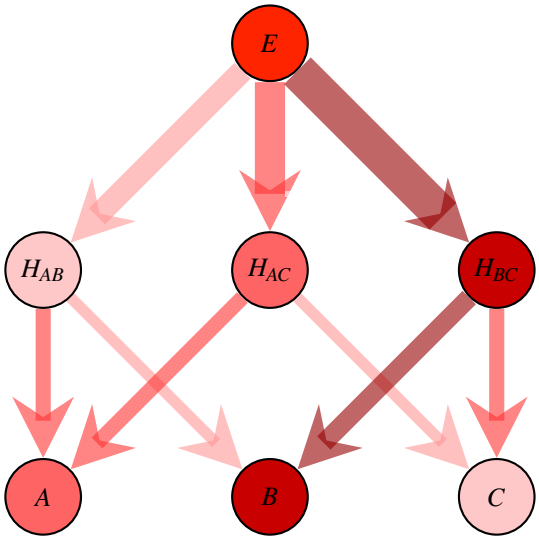


Figure 4.11: Relevance propagation for the same network as in fig. 4.11, with node colors indicating the amount of relevance (more intense red indicates higher relevance), and edge thickness proportional to the amount of relevance propagated along each edge. The relevance values for neurons of the hidden layer are $H_{AB} = 0.250$, $H_{AC} = 0.333$, $H_{BC} = 0.417$ and the relevance values trickled down to input layer nodes are $A = 0.333$, $B = 0.345$, $C = 0.321$. This captures the preference for B observed in Experiment 1 of chapter 3.

where:

- $(z_{ij})^+$ and $(z_{ij})^-$ are the positive and negative parts of z_{ij}, respectively.
- $(z_j)^+$ and $(z_j)^-$ are the sums of positive and negative contributions to z_j, respectively.
- α and β are parameters such that $\alpha - \beta = 1$.

This rule is essentially similar to the $\text{LRP}_{\alpha\beta}$ rule presented earlier for LFP. It partitions contributions into positive and negative components that are normalized separately and assigned different coefficients. The constraint $\alpha - \beta = 1$ ensures that total relevance is conserved across layers, i.e., the sum of relevance scores at any given layer equals the relevance assigned at the output layer. A common parameterization for the α and β coefficients, is $\alpha = 1$, $\beta = 0$. This forces relevance to flow **exclusively** through paths that yield a positive contribution to the output *in the actual world*.

This choice has a very valuable advantage, in that it automatically assign a relevance score of zero to variables that do not play any causal role in bringing about the outcome in the real-world of reference, such as $C = -1$ in the example above. This cumulates with another advantage implicit in relevance propagation rules in general (including the one in eq. (4.39)), which is that variables that are unconnected to the outcome in the relevant network will not inherit any relevance. What this means is that relevance propagation schemes dispense us with the necessity of adding a pre-selection step to our theory of causal selection judgments, whereby the relevant candidates for causal selection are filtered out. Current counterfactual theories of causal selection have to assume such a pre-selection, usually done on the basis of *separate* theories of actual

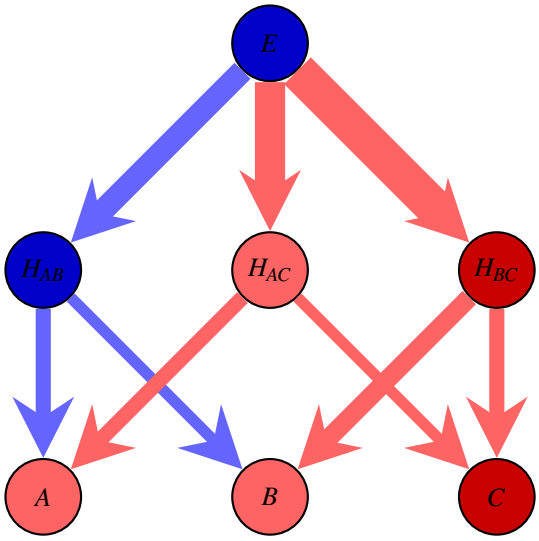


Figure 4.12: Relevance propagation in the negative activation example. Blue nodes and arrows represent negative relevance, while red nodes and arrows represent positive relevance. Note how C ends up with a high positive relevance score, despite being the one variable that does *not* contribute to the outcome.

causation (**halpern:2005**; such as e.g. Halpern 2015, 2016a). Using a LRP-measure in which signals only travel through positive paths takes care of such pre-filtering *de facto* without having to recruit separate a separate selection process. It unifies categorical notions of actual causation with graded notions of causal importance in a single measure.

For this reason we endorse the $\text{LRP}_{\alpha\beta}$ rule with those parameters. One nuance we add is that when the real-world outcome is negative we want relevance to flow only through negative activation paths, and block positive ones. This can be done by setting $\alpha = 0$ and thus $\beta = -1$ in those cases, which blocks positive paths while preserving the $\alpha - \beta = 1$ constraint.

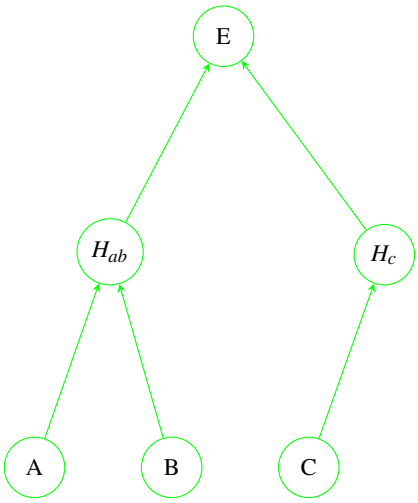
4.6.3 From Relevance Scores to Causal Impact Scores.

Desiderata. The relevance propagation procedure just defined allows us to track the contribution of each neuron to the outcome of interest. We do want the causal importance $\kappa(C, O)$ of a cause C (singular or plural) should be proportional to the relevance carried by causes $c \in C$, but not only that. The notion of causal importance that we are after is one such that the quality or intuitive attractiveness of an explanation for O that mentions C should be directly proportional to $\kappa(C, O)$. Yet for an explanation to be satisfying it is not enough that the facts it mentions be relevant, it should also be the case that the processes by which it influences the outcome be clear. For the sort of logical functions we are looking at presently, the relevant notions of processes are those charted by the neural paths, or equivalently the clause structure of the programs by which we represent those functions. Consider the simple program below, for illustration, which captures the causal rule $E := (A \wedge B) \vee C$:

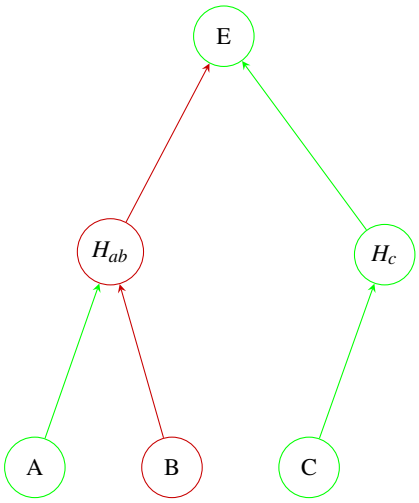
$$\{E \leftarrow A, B; \quad E \leftarrow C\} \tag{4.43}$$

Everything else being equal with respects to A, B, C's individual relevance scores, we want to say that, in a context where all events have occurred (as represented in fig. 4.13a) the plural cause $A \oplus B$ is a more attractive explanation than the plural $A \oplus C$, by virtue of the fact that the former brings about the outcome via a single program clause, while the latter "spreads" its contribution over two different clauses, corresponding to two separate procedures over that program. Interestingly, notions of actual causation such as Halpern (2015) do not automatically give us such groupings, as they would predict (in a situation where all A, B, C occurred) that the minimal sets of actual causes of the outcome for a rule like this one are $\{A, C\}$, $\{B, C\}$.

To highlight another desideratum, suppose that in some context A, C happened but B didn't (as pictured in fig. 4.13b). Here we would like to penalize any explanation that mentions A, no matter how correlated with B across counterfactuals it may otherwise be, because A cannot bring about E in the absence of B. Hence trying to derive E from A would result in unnecessary procedural steps, something we want to penalize.



(a) Network corresponding to the toy program (4.43), in the context where A = 1, B = 1, C = 1.



(b) Network corresponding to the toy program (4.43), in the context where A = 1, B = -1, C = 1.

Figure 4.13: Networks implementing the program in (4.43)

Hence we would like to construct a notion of *Path Complexity* $\mathscr{C}(C, O)$ for a certain causal explanation C of the outcome O with which we can turn our Relevance score assessments into causal impact scores via:

$$\kappa(C, O) = \frac{\sum_{c \in C} R_c}{\mathscr{C}(C, O)}, \tag{4.44}$$

It acts as a denominator for the sum of relevance contributions held by causes in C, whose point is to reward those causes whose contribution to the outcome is most straightforward. Intuitively, we would like that notion of complexity to track down the number of "routes" through which the signal from causes to the outcome travels in the network, capturing the intuition that A and B, in the context of the toy program in 4.43, affect E via the same route, which is different from the route by which C affects E.

Introducing Routes and Active Routes. The notion that we are after bears some similarity with the notion of an *active route* introduced by Hitchcock 2001a, 2007 in the context of a theory of *actual causation*. The

notion of active route seeks to capture the precise chain of variables or events through which a cause influences an outcome in a directed graphical model (often formalized via structural equations). It can be captured in classical SCMs with the definitions below.

Definition 3 (Route)**.** *Consider a directed acyclic graph (DAG) whose vertices represent variables (or events), with edges depicting direct influences. A* route *from a vertex C to a vertex O is a finite sequence of vertices*

$$(C = v_0, v_1, \ldots, v_{k-1}, v_k = O)$$

such that each (v_i, v_{i+1}) *forms a directed edge in the DAG.*

Definition 4 (Active Route)**.** *A route*

$$(C = v_0, v_1, \ldots, v_k = O)$$

is said to be active *in a particular scenario if, when all off-route variables (i.e. those not on the path* $v_0 \to \cdots \to v_k$*) are held fixed at their actual values, a suitable change in C (from its actual value to some alternative) brings about a change in O. In other words, O* counterfactually depends *on C once we restrict our attention to that route alone.*

The notion is close in spirit to the one we are seeking to capture, but the definition of active route here will not make enough sense once we step out of SCMs that can represent "true" boolean function. Because the networks we are dealing involve continuous weights and activations functions, we expect a change in the activation of any neuron to make some difference to the activation value of its children in the network no matter what. In the network represented in Figure 4.10, for example, this would lead us to say that even in the actual-world context where $C = -1$, there are still two active routes from C to E (one from H_{AC} and one from H_{BC}) because any change in the activation of C would make at least some change in the activation of those hidden nodes, which in turn would make some change in the activation of E.

One way to rescue it could be to put stronger requirements on the kind of difference-making required to satisfy the conditions in definition 4, by using the threshold value A_{min} to "binarize" the interpretation of activation values of successive neurons on the path, and then ask for example ask that each neuron activation on the path be part of a minimally sufficient condition for putting the activation of the next neuron above A_{min} or below $-A_{min}$, using a strategy similar to Baumgartner (2008, 2013)'s refinement of Hitchcock's criterion, defined schematically below, but replacing the notion that a variable takes its actual value with the notion that its activation is above or below a certain threshold.

Definition 5 (Baumgartner-Active Route, schematic)**.** *Let* $\langle V, E \rangle$ *be a causal model with variables V and structural equations E, represented by a directed acyclic graph G. Let C and O be variables in V. A route*

$$(C = v_0, v_1, \ldots, v_k = O)$$

is Baumgartner-active *if, for each* $i < k$*, there exists a* minimal *set of variables* $S_i \subseteq V$ *with* $v_i \in S_i$ *such that* S_i *is a minimally sufficient condition for* v_{i+1} *to take its actual value. Formally:*

(i) S_i *is sufficient for* $v_{i+1} = a_{v_{i+1}}$ *in the actual situation,*

(ii) S_i *is minimal in the sense that no proper subset of* S_i *is likewise sufficient.*

Though I am sympathetic to the spirit of this notion too, it will not be flexible enough to apply to the cases I am interested in. This is true for at least two reasons. One is that, because I propose that causal

importance scores are computed over a network whose weights have been updated (by the process described in Section 4.5), there is no guarantee that the relation between neuron activations in the updated network $\mathsf{N}^{(\mathrm{up})}$ maintains the relation to the A_{min} that they have in the network N coming fresh out of CILP translation. Secondly, as I will explain in more details in Chapter 7 of this dissertation, part of the point of defining measures of explanation over neural network is to make it possible to build and interpret such explanation over networks that are still in the process of figuring out the right causal theory, and whose yet "immature" weights therefore can't be expected to respect the sort of relationship entailed by this kind of thresholding criterion.

Positive Activation Paths For all of these reasons, I propose the following definition. It takes advantage of the fact that, in the sort of networks we are dealing with here, bringing about a certain outcome is always about positively activating the corresponding neuron (I'll explain later in this section how that is true even for negative outcomes) to capture the notion of "routes by which causes influence an outcome" in terms of the signs of connection weights along the network route that connects them to the outcome.

Definition 6 (Positive activation paths in a neural network)**.** *Let*

$$\mathsf{N} = \langle \mathsf{V}, \mathsf{E}, \mathbf{W}, \boldsymbol{\Theta} \rangle$$

be a neural network composed of a set of neurons V*, layers* V_{in}*,* $\mathsf{V}_{\text{hidden}}$*,* V_{out}*, directed edges* $\mathsf{E} \subseteq \mathsf{V} \times \mathsf{V}$*, weights* $\mathbf{W} = \{w_{ij} \mid (i,j) \in \mathsf{E}\}$*, and biases* $\boldsymbol{\Theta} = \{\theta_j \mid j \in \mathsf{V}\}$*. Let AW be a state of the network, obtained by assigning activation values* a_{in}^{AW} *to every neuron* $v \in \mathsf{V}_{in}$ *followed by a forward pass through the next layers to get an activation value* a_v^{AW} *for every* $v \in \mathsf{V}$*. Let* $C \subseteq \mathsf{V}$ *be a set of candidate cause neurons and* $O \in \mathsf{V}_{\text{out}}$ *be an outcome neuron. Let* $Y \in \mathsf{V}$ *be a neuron that directly connects to O with an edge* $(Y,O) \in \mathsf{E}$*. (Note that in the simplest case of three-layered neural network as those looked at so far, every hidden neuron* $h \in \mathsf{V}_{hidden}$ *can be one such Y).*

A ***positive activation path*** *from the set C to the neuron O in AW is a directed path*

$$p = (\, c = v_0, v_1, \ldots, v_{k-1} = Y, v_k = O\,) \quad \textit{with } c \in C,$$

such that:

1. $(v_i, v_{i+1}) \in \mathsf{E}$ *for all* $0 \leq i < k$ *(i.e. there is an uninterrupted directed path from c to O in the network),*
2. *The product of all edgewise contributions along p is also strictly positive, i.e.*

$$\prod_{i=0}^{k-1} \left(W_{v_i \to v_{i+1}} \cdot a_{v_i}^{AW} \right) \; > \; 0,$$

3. *The input contribution* $z_{Y \to O}$ *of Y, the last neuron on the path to O, is also strictly positive in AW:*

$$z_{Y \to O} = w_{Y \to O} \cdot a_Y^{AW} > 0.$$

The second condition in this definition ensures that C makes a positive contribution to O through the path p in AW. In the context represented in fig. 4.13b, for example, it captures the fact that C contribute to E but A doesn't, because the hidden neuron by which A's positive input could have reached E is not active (because of B), so that A's contributed is "negated" at the next step on the path to E. The third condition also ensures that

B does not mistakenly get counted as being on a positive activation path, just out of the fact that the product of contributions on its path to E happens to be positive.

This notion of a positive action path will serve as the basic building block by which we'll be able to build a notion of path complexity. Now it is not enough on itself to do so, as for example it does not handle the fact that, e.g. in fig. 4.13a, A and B both connect to E via the same path. To have a notion of complexity that cares about the number of processes or mechanisms by which a certain cause exerts its effects, we'll need to restrict penalties for complexity to path that share no edge in common. I propose to do so via the notion of *edge-disjoint paths*, as defined below.

Definition 7 (Edge-Disjoint Paths)**.** *Let $p = (v_0, v_1, \ldots, v_k)$ and $q = (u_0, u_1, \ldots, u_m)$ be two directed paths in a network $\mathsf{N} = \langle \mathsf{V}, \mathsf{E}, \mathbf{W}, \boldsymbol{\Theta} \rangle$, where each consecutive pair of nodes in p or q appears as a directed edge in E. We say that p and q are* edge-disjoint *if, for every $i \in \{0, \ldots, k-1\}$ and every $j \in \{0, \ldots, m-1\}$,*

$$(v_i, v_{i+1}) \neq (u_j, u_{j+1}).$$

That is, the two paths do not share any directed edge in E.

Counting edge-disjoint paths. On the basis of Definition 6, let's now denote

$$\mathbf{P}^+(C, O) = \big\{ p \mid p \text{ is a positive activation path from } C \text{ to } O \big\} \tag{4.45}$$

the set of all positive activation paths from C to O (implicitly defined with respect to a network N and actual-world AW). A subset $\mathsf{D} \subseteq \mathbf{P}^+(C, O)$, is said to be *edge-disjoint* if every pair of distinct paths in D is edge-disjoint in the sense of Definition 7. For convenience, we note the set of all such subsets:

$$\mathbf{D}^+(C, O) = \Big\{ \mathsf{D} \subseteq \mathbf{P}^+(C, O) \;\Big|\; \text{all paths in } \mathsf{D} \text{ are edge-disjoint} \Big\} \tag{4.46}$$

which is the *family of all edge-disjoint subsets of positive activation paths* from C to O. A natural way to capture the *total number* of edge-disjoint positive paths from C to O is to take the maximum cardinality of any edge-disjoint subset in that family. Concretely, this defines:

$$\#_{\text{ED}}(C, O) = \max_{\mathsf{D} \in \mathbf{D}^+(C,O)} \big|\mathsf{D}\big|, \tag{4.47}$$

In words, $\#_{\text{ED}}(C, O)$ is the maximum number of positive activation paths one can choose from C to O so that no two chosen paths share any directed edge. Not that the number of neurons that directly connect and send positive activation into O in the relevant network puts an upper bound on $\#_{\text{ED}}(C, O)$, so that in the sort of simple networks we look at here, computing $\#_{\text{ED}}(C, O)$ can be done simply by counting the number of such neurons to which C directly contribute (i.e. the number of activated hidden neurons in networks like those in Figure 4.13).

Finally, let us define the set of *idle causes* in C that do not participate in any positive path to O as:

$$\text{Idle}(C, O) = \big\{ c \in C \mid \forall\, p \in \mathbf{P}^+(C, O), c \notin p \big\}. \tag{4.48}$$

On these grounds, we can define a **path complexity** measure as follows:

$$\mathscr{C}(C, O) = \#_{\text{ED}}(C, O) + \big|\text{Idle}(C, O)\big|, \tag{4.49}$$

where $\#_{\text{ED}}(C,O)$ counts the number of distinct paths by which C contributes to O and $|\text{Idle}(C,O)|$ counts the number of causes in C that do not contribute to O via any path. Intuitively, a cause C constitutes a straightforward explanation to the extent that it minimizes the sum of these two quantities. This is integrated in our measure of causal importance by making $\mathscr{C}(C,O)$ the denominator in Equation (4.44)
Note that one could in principle want to assign different penalties to each of these two factors, by assigning coefficient to the terms of the equation above:

$$\mathscr{C}(C,O) \;=\; \beta_1 \#_{\text{ED}}(C,O) \;+\; \beta_2 \big|\text{Idle}(C,O)\big|, \tag{4.50}$$

with $\beta_1, \beta_2 > 0$. I don't exploit this degree of freedom here however, and stick to a default parameterization where $\beta_1 = \beta_2 = 1$.

4.6.4 Applying the measure to concrete examples: positive outcomes

In this final section, I propose to illustrate the workings of the causal importance I just define on some concrete cases. I'll focus in particular on cases in question will be taken from our Experiment 2 on plural causes presented in Chapter 3. Recall that the rule of the game people played in that experiment said that to win a round of the game, one must get a colored from both urns A and B, or from both urns C and D. My assumption throughout is that this rule is captured by the general logic program:

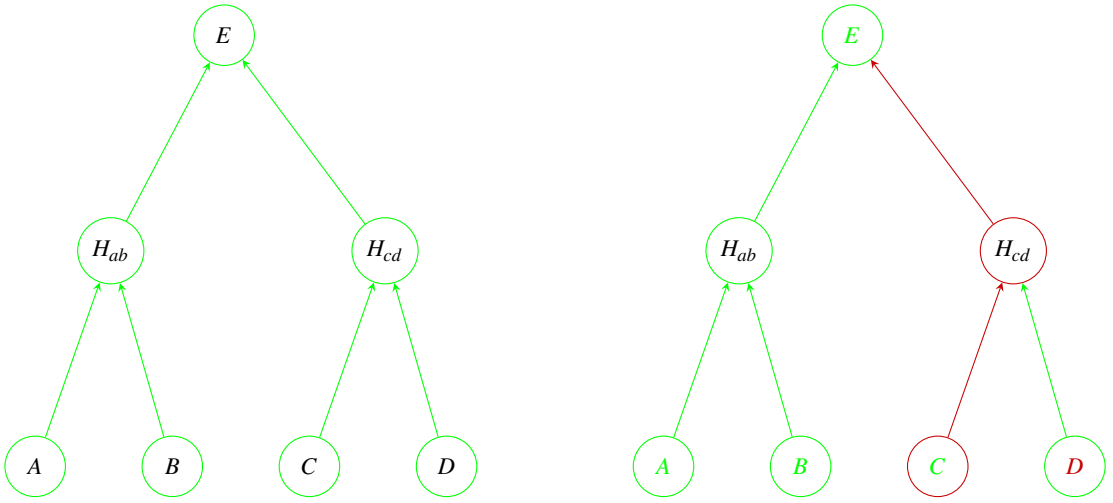
$$\mathsf{LP}_{\mathit{urns}} = \{E \leftarrow A \wedge B\,; \quad E \leftarrow C \wedge D\}. \tag{4.51}$$

I first briefly present the application of the measure to the two experimental conditions where the outcome to explain was positive (the player won a round of the game), the OVERDETERMINED POSITIVE and TRIPLE 1 conditions, corresponding the network representations in Figure 4.14. I then turn the OVERDETERMINED NEGATIVE and TRIPLE 0 rounds where subjects where tasked with explaining negative outcomes $\neg E$ ("The player lost a round of the game"). Because classical negation cannot be represented by default in the program in (4.51) and the corresponding networks, I introduce a closed-world operator to handle explanations of negations. It involves the introduction of a new path $\neg E \leftarrow \sim E$ (different from the extended program clause with $\sim E$ as body and $\neg E$ as head). I explore how the consequences this has on the way causal importance is computed account for the patterns of judgments observed in the negative conditions of our experiment. To enhance clarity, the presentation made here stays very schematic and concentrates on high-level patterns. Readers interested in more precise numerical/modeling details can find them in the code attached with this dissertation.

Positive conditions. For positive conditions, the complexity measure presented in eq. (4.49) will capture the following patterns.

- In the OVERDETERMINED POSITIVE condition (fig. 4.14a), the plural cause explanations $A \oplus B$, $C \oplus D$ each connect to E via one and the same path. As a result, we expect each of them to always be more attractive than the singulars that they contain. This is because an explanation like $A \oplus B$ will sum the relevance accumulated over both A and B, without increasing the denominator $\mathscr{C}(C,O)$, compared to the explanation that cites only one of these causes alone.

  The same logic also explains why cross-paths explanations like $B \oplus C$ are unattractive, even if they gather individual variables with high relevance, and why triples like $A \oplus B \oplus C$ get penalized though not *more* than $B \oplus C$. A triple like $A \oplus B \oplus C$ incurs an additional penalty compared to $A \oplus B$ (because it



(a) Overdetermined positive condition

(b) TRIPLE 1 condition

Figure 4.14: Neural networks corresponding to LP_{urns}. In the OVERDETERMINED POSITIVE condition, the player draws a colored ball from each of the four urns and wins the round of the game. In the TRIPLE 1 condition, the player draws a colored ball from all of the urns except for C, and also wins. The figures above schematically represent the relevant network activation. Nodes' colors represent the activation signs of individual neurons, and edge colors represent the sign of the input sent along that connection. (Green for positive activations/inputs, Red for negative ones).

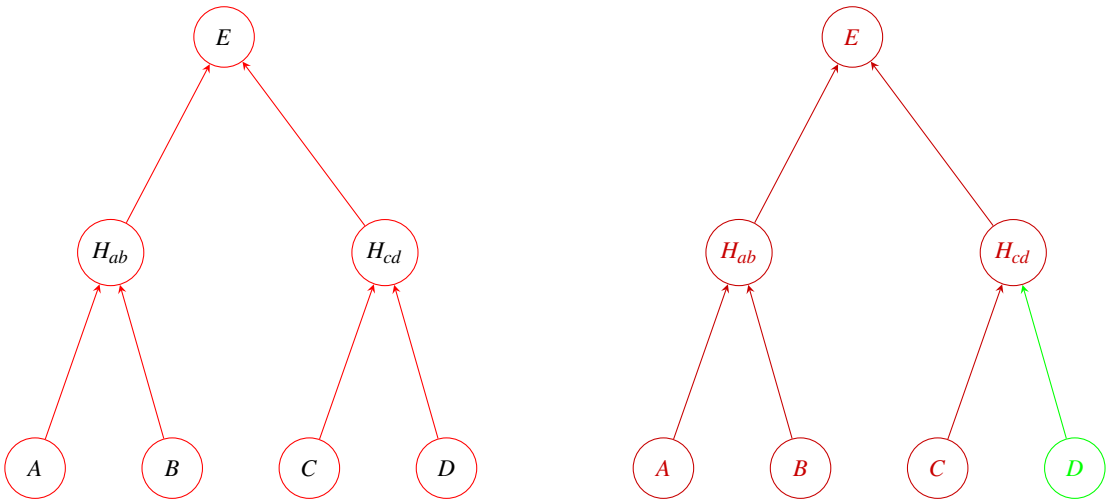
involves one more path), but it does not add more paths than $B \oplus C$ already does, hence why we expect it to be systematically higher regardless of the specific relevance scores of the variables it contains. Such examples illustrate the way in which the $\mathscr{C}(C,O)$ penalty is not about the number of variables but about the number of paths they take to the outcome.

- In the TRIPLE 1 condition, the same measure nicely accounts for the fact (verified in our data) that plural explanations that involved the idle variable D and then some other variables (such as $A \oplus D$ or $A \oplus B \oplus D$) were systematically rated with a score that was directly intermediate between the same explanation without D (e.g. A or $A \oplus B$) and zero. This follows from the fact, that D, being an idle variable in this context, will add zero relevance to whichever plural causes to which it is appended, but will increase the denominator $\mathscr{C}(C,O)$ by incrementing it with one idle variable. Explanations that mentioned just D where on the other hand given ratings very close to zero, which the same measure also predicts.

  This highlights another advantage of the measure as compared to usual accounts that handle idle variables with categorical exclusions (based on theories of actual causation), which is that it can handle the *incremental* (as opposed to categorical) nature of unattractiveness attached to such explanations when the variables in question are embedded in plurals. It also relates to one of the defining features of plurals (as opposed to classical conjunctions) in natural language more generally, which is their *tolerance to exceptions*[2] .

[2]It is worth remarking that tolerance to exceptions in natural language plurals is itself also parametrized to some notion of relevance,

4.6.5 Applying the measure to concrete cases: negative outcomes



(a) Overdetermined negative condition

(b) TRIPLE 0 condition.

Figure 4.15: Neural networks corresponding LP_{urns}, in the negative conditions.

The measure of causal importance defined above will require a different handling in situations where a subjects has to explained negative outcomes. To explain why a player *loss* a round of the game amounts to providing an explanation for $\neg E$. But the network $\mathsf{N}_{\mathsf{LP}_{urns}}$ (or any other network that is understood as the translation of a general logic program) is not equipped to derive a negative outcome like $\neg E$. Network states where $a_E = -1$, as those represent in Figure 4.15 cannot be understood as states where a certain configuration of inputs led to a loss, only to states where that configuration failed to yield an occurrence of E.

Subjects may feel safe to assume that their internal generative model tracks all of the facts relevant to the occurrence of the outcome E (especially so in a simple setting like that of this experiment), so that from the fact that a configuration of draws fails to make a player win they might derive that it makes the player lose ($\neg E$). But this does not mean that their internal model encodes this assumption by default. This is analogous to mental model theorists's observation that people grasp the meaning of a proposition by focusing on models that make that proposition true – while have to engage extra cognitive work to represent the conditions that would falsify it. Here we assume that people's internal programs are program for generating occurrences of an outcome E, and are not *ipso facto* endowed it with the power to track down occurrences of $\neg E$.

For them to generate and explain such negative outcomes, people have to engage in an additional procedural step, whereby they allow themselves to derive $\neg E$ from a failure to derive E ($\sim E$)[3]. I will assume

as remarked by e.g. Malamud (2012). For example, a sentence like: "Mary read the books on her reading list" is acceptable even in case Mary only read a few of the books on her reading list, if the context makes it such that the subset of books she read are particularly relevant to some contextual purpose (e.g. Mary's incoming exam).

[3]Note that nothing about this assumption is in contradiction with the fact that determining which outcome counts as the positive outcome E, and which counts as its negation $\neg E$ is entirely framing dependent. In other words, had we instructed our experimental subjects in the conditions for *losing* a round of the game, instead of the winning conditions, they would have build a program that focuses on deriving losing events like $\mathsf{LP}_{\text{loss}} = \{L \leftarrow\sim A, \sim C\,; \quad L \leftarrow\sim A, \sim D; \quad L \leftarrow\sim B, \sim C; \quad L \leftarrow\sim B, \sim D\}$ (where the atom L tracks losing events), or even, more simply: $P_{\text{loss}'} = \{F \leftarrow A, C\,; \quad F \leftarrow A, D; \quad F \leftarrow B, C; \quad F \leftarrow B, D\}$. (where each atom A, B, C, D is taken to represent the event of drawing a white ball from the corresponding urn). Then, they would have needed to engage the sort of

that doing so for subjects involves the use of a new closed-world operator, equivalent to extended the program LP_{urns} with a clause

$$\neg E \leftarrow \sim E \tag{4.52}$$

that explicitly handles this derivation. This clause is not however handled in the standard way, as an extended program clause that CILP-translates into its own network path as represented in Figure 4.16, for several reasons. One reason is that we want to avoid allowing for extended program clauses by default as much as possible, as explained in Section 4.2; it that sense it is more suited to understand the assumption in eq. (4.52) in terms of a special closed-world operator that people convoke only when they are explicitly tasked to explain negative outcomes, rather than part of the knowledge base that they represent by default. A second (and related) reason is that having a new node E feature in V_{in} blurs the distinction between causes and effects which we assume people have. It would suggest that, for example, the statement "The player lost because he didn't win" could be a valid explanation, on a par with, e.g. "The player lost because he drew a white balls from urns B and C", although we intuitively like to think of the first statement as a non-explanation.

For all of these reasons, I instead assume that the step represented by $\neg E \leftarrow \sim E$ is implemented in a non-standard way, via a direct network path from the output node E to another output node $\neg E$, as pictured in Figure 4.17. This means that the operator shifts the network structure is "shifted" to push back its outcome layer to $\mathsf{V}_{\text{out}'} \supseteq \neg E$ as $\neg E$ becomes the explanandum, so that the former output layer on which E is now located on a second hidden layer $\mathsf{V}_{\text{hid2}} \supseteq E$, the first hidden layer being the original $\mathsf{V}_{\text{hid}} \supseteq \{H_{ab}, H_{cd}\}$.

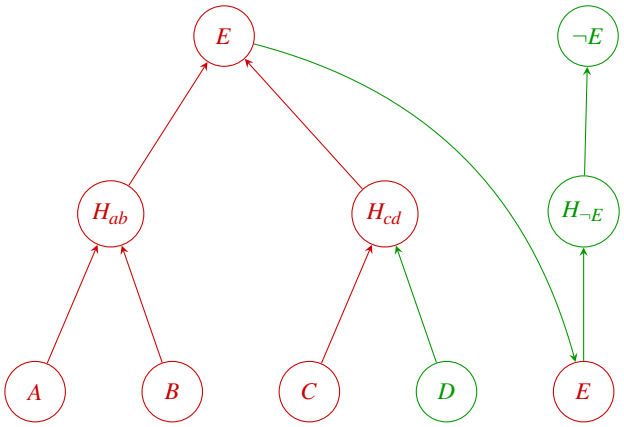


Figure 4.16: Extended network $\mathsf{N}_{\mathsf{LP}_{urns}}$, with node and edge activation matching the TRIPLE 0 condition. The weight on the edge from $E \in \mathsf{V}_{\text{out}}$ to $E \in \mathsf{V}_{\text{in}}$ has a weight of 1 and simply transmits its input to the node in E whose activation follows the standard linear function $g(x) = x$. The weight from $E \in \mathsf{V}_{\text{in}}$ is negative, so that $H_{\neg E}$ is activated when the activation of $E \in \mathsf{V}_{\text{in}}$ is negative.

Flattening of path complexity for negative outcomes. One direct consequence of adding the edge $E \rightarrow \neg E$ is that it *flattens* the path complexity of every set of causes $C \subseteq \mathscr{P}(\mathsf{V}_{\text{in}})$. This is because *all* of the positive activation paths from any combination of inputs A, B, C to $\neg E \subseteq \mathsf{V}_{\text{out}'}$ in the network in fig. 4.17 all share an edge in common which is $E \rightarrow \neg E$. As a result, there is no more penalty to the explanation $A \oplus B \oplus C$ ("John lost because he drew white balls from urns A, B, and C"), than there is to shorter explanations $B \oplus C$

closed-operator described in this section to explain positive outcomes.

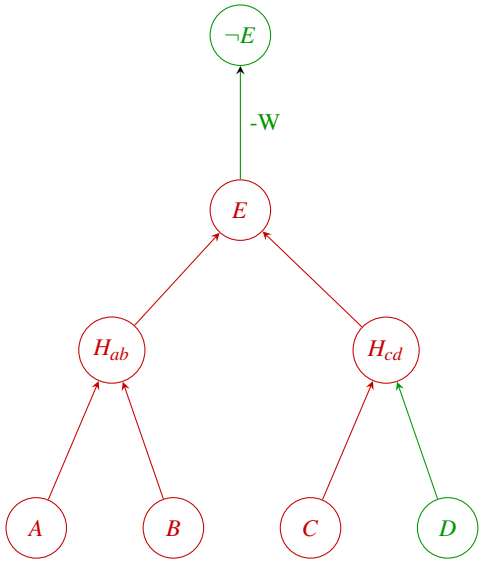


Figure 4.17: Network $\mathsf{N}_{\mathsf{LP}_{urns}}$, here with node and edge activation matching the TRIPLE 0 condition. The weight $w_{E\to\neg E}$ is negative, which, in a context like this where $E<0$, leads the activation input $z_{E\to\neg E}$ to be positive. The closed-world operator shifts the network structure is "shifted" to push back its outcome layer to $\mathsf{V}_{\text{out}'} \supseteq \neg E$ as $\neg E$ becomes the explanandum, so that the former output layer on which E is now located on a second hidden layer $\mathsf{V}_{\text{hid2}} \supseteq E$, the first hidden layer being the original $\mathsf{V}_{\text{hid}} \supseteq \{H_{ab}, H_{cd}\}$

or even just B. More volubile explanations are thereby implicitly encouraged, because they gather more relevance at no extra cost. This captures one half of the surprising patterns revealed in the losing rounds of our experiment, which is the fact that (contrary to what we observed in the winning rounds) people did not penalize explanations that were redundant for explaining the loss. On the contrary, subjects favored explanations that contained as many variables as possible. Pairs were systematically rated higher than any of the variables they contained, and triples systematically higher than pairs. The only plurals that $\mathscr{C}(C,O)$ would penalize are the plurals that mention the variable D in the TRIPLE 0 condition (e.g. $B \oplus D$) because D would classify as an idle cause in that particular context. This would lead these plurals to be penalized in the same way as plurals containing D in the TRIPLE 1 condition. Although we did not test subjects on these judgments, the prediction seem intuitive.

Negative outcomes reversal abnormal inflation, but not abnormal deflation. The other surprising pattern of judgment that we observed in the losing rounds of our experiment. In losing rounds, people did not prefer to explain the loss by the fact that the play drew white balls from urns B and C, which were the urns containing the highest proportion of white balls. Yet this is what we wold expect them to do if they were tracking the sufficient conditions for losing, by the logic of abnormal deflation. Instead they preferred explanations that mentioned urns A and D, the urns that contained the smallest proportion of white balls.

I propose to account for this pattern as a special case of a more general fact about the way in which different causal structures interact with negative outcomes. Specifically, I explain why we expect *abnormal inflation* (with respects to input-to-hidden nodes) to be reversed when it comes to explaining negative outcomes, but not *abnormal deflation*. That this is so is independently evidenced in other experiments that

did not look at plural cause judgment such as Gerstenberg and Icard (2020) where participants' judgments for negative outcomes in the *disjunctive* cases where $E := A \vee B$ were neatly captured by treating negative outcomes as the classical negation $\neg E := \neg A \wedge \neg B$. But the same strategy did not work in the conjunctive case where $E := A \wedge B$. My explanation hinges on the way in which the weight update mechanisms presented in Section 4.5 interact with the extension of the network with the new outcome node $\neg E$. For clarity I focus on simpler cases of settings where only two causes are involved, i.e. with programs $\mathsf{LP}_{\text{or}} = \{E \leftarrow A\,; \quad E \leftarrow B\}$ and $\mathsf{LP}_{\text{and}} = \{E \leftarrow A \wedge B\}$. Extrapolating that logic to the explain the results in our experiment with the causal setting $\mathsf{LP}_{\text{urns}} = \{E \leftarrow A, B\,; \quad E \leftarrow C, D\}$ is then straightforward. Readers interested in more precise modeling/numerical details can however find them in the code attached with this dissertation if desired.

No reversal of abnormal deflation. Consider the logic program:

$$\mathsf{LP}_{\text{or}} = \{E \leftarrow A\,; \quad E \leftarrow B\}$$

And the corresponding network $\mathsf{N}_{\mathsf{LP}_{\text{or}}}$, extended with the closed-world path $\neg E \leftarrow \sim E$, in the actual world of reference where neither A nor B occurred (and therefore E didn't occur either), as represented in Figure 4.18a. As explained above, the introduction of $\neg E \leftarrow \sim E$ means that subjects, as they perform weight-updates over their internal model across counterfactual scenarios that they consider, do so with $\neg E$, instead of E, taken as the explanandum. This means that the rewards relevant for weight updates will be propagated from the top node $\neg E$. Recall also from the previous Section 4.5 that for each counterfactual state $s^{(i)} \in \{s^{(0)}, \ldots, s^{(n)}\}$ considered by the subject, the reward $r^{(i)}_{\neg E}$ associated with the outcome $\neg E$ satisfies (by eq. (4.36))

$$r^{(i)}_{\neg E} = \mathbf{y}_{\neg E} \cdot a^{(i)}_{\neg E},$$

And that this reward score is propagated to previous layers via eq. (4.37), reproduced below:

$$r'_{ij} = \begin{cases} \dfrac{|z^+|}{|z^+_j| + |z^-_j|} \cdot \dfrac{z_{ij}}{z^+_j} \cdot \text{sign}(z_j) \cdot r_j & \text{if } z_{ij} \geq 0, \\ \dfrac{|z^-|}{|z^+_j| + |z^-_j|} \cdot \dfrac{z_{ij}}{z^-_j} \cdot \text{sign}(z_j) \cdot r_j & \text{otherwise.} \end{cases} \tag{4.53}$$

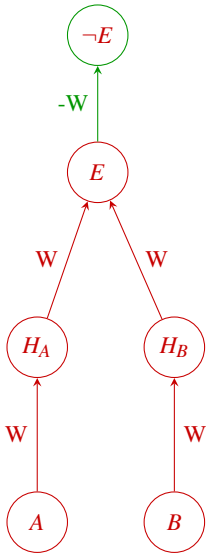
Given that $\mathbf{y}_{\neg E} = 1$ (the outcome was observed to be $\neg E$ in the world of reference), this simplifies to

$$r^{(i)}_{\neg E} = a^{(i)}_{\neg E} \text{ across states } s^{(i)} \in \{s^{(0)}, \ldots, s^{(n)}\} \tag{4.54}$$

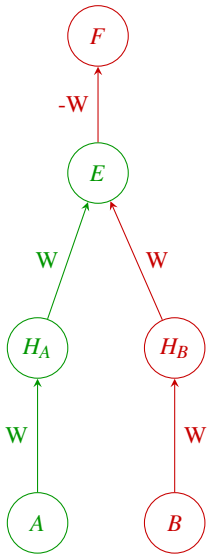
A direct consequence of that, by eq. (4.53), and using the same logic by which we showed in section 4.5 that the reward values of hidden nodes are always positive (even when their activation is negative) when E is the target outcome and $\mathbf{y}_E = 1$ is:

$$r^{(i)}_{E} = |a^{(i)}_{\neg E}| \text{ across states } s^{(i)} \in \{s^{(0)}, \ldots, s^{(n)}\} \tag{4.55}$$

i.e. the reward value on E is always positive and equal to the activation value of $\neg E$. This highlights how E behaves as a hidden neuron sending direct input into $\neg E$ just like neurons $h \in \mathsf{V}_{\text{hidden}}$ were to E in positive cases, when E was the explanandum. Recall that asymmetries between different weights in $W_{\text{hidden} \to E}$ are introduced by the weight update process in "contrastive" states where neurons $h \in \mathsf{V}_{\text{hidden}}$ do not all have the same activation states. If all of them are active, or all of them are inactive, then they would be reward



(a) Network $\mathsf{N}_{\mathsf{LP}_{\text{or}}}$ in a world of reference where $a_A^{AW} = a_B^{AW} = -1$, extended with the path $\neg E \leftarrow \sim E$.



(b) Counterfactual state in which $a_A = 1$ and as a result $a_E \approx 1$.

Figure 4.18: Illustration of $\mathsf{N}_{\mathsf{LP}_{\text{or}}}$ and a contrast case scenario.

equally, so no asymmetry between them is introduced as a result. So for the purpose of tracking down the high-level dynamics of the network they can be ignored. Let's call $SC \subseteq \{s^{(0)}, \ldots, s^{(n)}\}$ the remaining subset of the relevant contrastive states, formally:

$$SC = \left\{ s^{(j)} \in \{s^{(0)}, \ldots, s^{(n)}\} \mid \exists h, h' \in \mathsf{V}_{\text{hidden}} \left(h \neq h' \wedge \text{sign}(a_h^{(j)}) \neq \text{sign}(a_{h'}^{(j)}) \right) \right\}. \tag{4.56}$$

One such state is illustrated in fig. 4.18b for the network $\mathsf{N}_{\mathsf{LP}_{\text{or}}}$. Recall also that in the networks built out the CILP procedure (and as also pictured in fig. 4.18b), the net input activation input z_E of a neuron $E \in \mathsf{V}_{\text{out}}$ (which is now $E \in \mathsf{V}_{\text{hidden2}}$ in the extended network, but all relations with anterior layers are preserved) is positive whenever one of the hidden neurons in $\mathsf{V}_{\text{hidden}}$ connected to it is positively activated, and negative otherwise. This means that

$$z_E^{(s)} > 0, \; \forall s \in SC, \tag{4.57}$$

As a result, and by virtue of eq. (4.55) above and the LFP update equations (4.31)-(4.35), reproduced below:

$$\delta_{w_{ij}}^{\text{lfp}} = \frac{|w_{ij}| \cdot a_i}{|z_j^-| + |z_j^+|} \cdot \text{sign}(z_j) \cdot r_j. \tag{4.58}$$

$$w_{ij}^{\text{new}} = w_{ij}^{\text{old}} + \eta \cdot \delta_{w_{ij}}^{\text{lfp}} \tag{4.59}$$

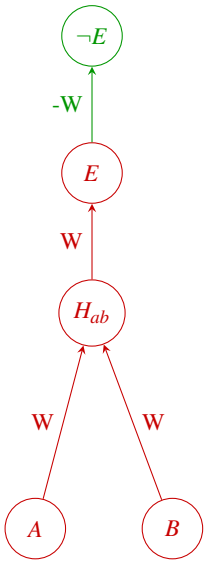
In all such contrastive cases $s \in SC$, the weights $w_{h \to E}$ from neurons $h \in \mathsf{V}_{\text{hidden}}$ that are positively activated (like H_A in fig. 4.18b) will get positive updates, while weights $w_{h' \to E}$ from neurons $h' \in \mathsf{V}_{\text{hidden}}$ that are negatively activated (like H_B in fig. 4.18b) will get negative updates. This dynamic ultimately rewards the weights in $w_{h \to E} \in W_{\text{hidden} \to \text{hidden2}}$ that originate from the $h \in \mathsf{V}_{\text{hidden}}$ most frequently activated across sampled

states, just like those were the most rewarded for positive outcomes. The ensuing relevance propagation process then leads to giving more importance to the most normal input causes that connect to them. This captures the non-reversal of judgments for negated disjunction documented in Gerstenberg and Icard (2020).

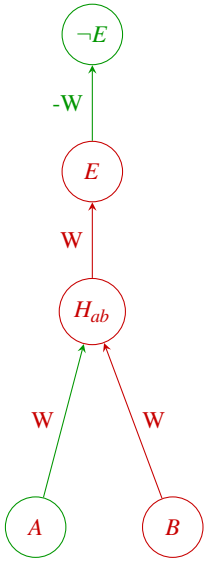
Reversal of abnormal inflation. Consider now the logic program:

$$\mathsf{LP}_{\text{and}} = \{E \leftarrow A \wedge B\}$$

And the corresponding network $\mathsf{N}_{\mathsf{LP}_{\text{and}}}$, extended with the closed-world path $\neg E \leftarrow \sim E$, in the actual world of reference where neither A nor B occurred (and therefore E didn't occur either), as represented in Figure 4.19a. As explained above, the introduction of $\neg E \leftarrow \sim E$ means that subjects, as they perform weight-updates over their internal model across counterfactual scenarios that they consider, propagate rewards from $\neg E$ instead of E. Since here as in the previous case $\mathbf{y}_{\neg E} = 1$, we have, by eq. (4.54) and eq. (4.55):



(a) Network $\mathsf{N}_{\mathsf{LP}_{\text{and}}}$ in the world of reference where $A = B = -1$, hence $E \approx -1$ and $\neg E \approx 1$



(b) Contrastive case with $A = 1$. Since the activation of H_{ab} is such that it needs both A and B to be on, $A = 1$ alone is not sufficient, hence H_{ab}, E stay negative and $\neg E$ stays positive.

Figure 4.19: AND-network $\mathsf{N}_{\mathsf{LP}_{\text{and}}}$ and a contrastive activation pattern. (a) Baseline state with inactive inputs. (b) Contrastive state demonstrating E's activation when both antecedents are present.

$$r_E^{(i)} = |a_{\neg E}^{(i)}| \text{ across states } s^{(i)} \in \{s^{(0)}, \ldots, s^{(n)}\}$$

A follow-up consequence, by eq. (4.53), is that the sign of the reward on hidden nodes will always be equal to that of z_E for every $h \in \mathsf{V}_{\text{hidden}}$:

$$\text{sign}(r_h^{(i)}) = \text{sign}(z_E^{(i)}), \quad \forall h \in \mathsf{V}_{\text{hidden}}. \tag{4.60}$$

Now, let's define a notion of contrastive states for *input nodes*, analogous to how we defined contrastive states for hidden neurons in eq. (4.56) as those network states where two hidden neurons with different activation signs are found. There are two key differences however. First, this notion applies specifically to a specific hidden neuron $h \in \mathsf{V}_{\text{hidden}}$, tracking states in which its *input predecessors* exhibit conflicting signals. This is because we might want to consider networks with several hidden neurons, each of which receives inputs from different sets of input neurons. Second, since unlike hidden nodes input nodes may connect through both positive and negative weights (if the source program contains clauses with negation-as-failure like $E \leftarrow\sim I$), we need to consider the sign of the product $z_{i,h}$ between an input neurons' activation a_i and the connecting weight $w_{i\to h}$ in a state $s^{(i)}$, which is given by:

$$\text{sign}(z_{i,h})^{(i)} := \text{sign}\left(a_I^{(i)} \cdot w_{i\to h}\right) \tag{4.61}$$

This defines the contrastive input states SCI_h relative to a neuron $h \in \mathsf{V}_{\text{hidden}}$ as those states where at least two of h's input predecessors send activation inputs with different signs:

$$SCI_h = \left\{ s^{(k)} \in \{s^{(0)}, \ldots, s^{(n)}\} \mid \exists i, i' \in \mathsf{V}_{\text{in}} \begin{pmatrix} i \neq i' \wedge (i,h) \in \mathsf{E} \wedge (i',h) \in \mathsf{E} \\ \wedge\ \text{sign}(z_{i,h})^{(k)} \neq \text{sign}(z_{i',h})^{(k)} \end{pmatrix} \right\}. \tag{4.62}$$

Figure 4.19b illustrates one such case for H_{ab} in $\mathsf{N}_{\mathsf{LP}_{\text{and}}}$. Now, just like we noted previously that in every contrastive state for $W_{\text{hidden}\to E}$ weights the sign of E will be positive, because contrastive state entail that at least one $h \in \mathsf{V}_{\text{hidden}}$ is positively activated, we note that in every contrastive state for $W_{\text{in}\to h}$ the sign of h will be negative, because these entail that at least one input neuron sends negative input signals to h. As a result, and by virtue of eq. (4.60) and the weight update equations (4.58)-(4.59), the weight $w_{i\to h}$ from each input predecessor i to h will be incremented with an update of the same sign as the a_i, as long as $z_E < 0$ (and with the same sign as $-a_i$ otherwise). The condition $z_E < 0$ is always satisfied whenever there is just one hidden neuron $h \in \mathsf{V}_{\text{hidden}}$ (as in Figure 4.19), and almost always in other cases as well, by virtue of the fact that sampling is initialized at at state where $z_E < 0$ (or this wouldn't be a negative outcome) and generates states very similar to the initial state for small sample sizes[4]. To increment each $w_{i\to h}$ with an update of the same sign as the a_i means to increase the absolute value of weights that positively contribute to z_h in $s^{(k)}$, and decrease the absolute value of weights that contribute negatively, which is the inverse of the pattern expected when the outcome is positive.

[4]Add to this the fact that for z_E to flip to a positive value is even less likely in the subset of states that we consider here where one of the hidden nodes is presumed to be "off" (or the state wouldn't be contrastive for any h).

4.7 Summary and intermediate conclusions

Before closing this chapter and the first part of this dissertation, a few remarks are in order. In the previous sections, I developed an account of causal selection judgments, grounded in the premise that people generate explanations through manipulations of an internal event-simulation mechanism. This account was guided by the basic notion that for an event C to cause an event E, C must be part of a generative process that brings about E. And that similarly for C to *be seen* as causing E, some mental tokening of C must take part in an *internal* generative process that yields a mental tokening of E. This intuition is arguably even more fundamental than the one that cashes out causality in terms of the counterfactual dependence patterns that can be represented in a SCM. It is because generative processes have a directional dataflow, which includes precursor relations between each step and the next, that they yield the counterfactual dependence patterns we are accustomed to see between causes and effects: if you disrupt one of the earlier steps in the internal procedure that yields E, you prevent it from producing E.

Focusing on processes offers a new perspective on causal explanations once we acknowledge that these processes must be implemented in a concrete device. Such implementation comes with various constraints that a model centered solely on describing the resulting counterfactual dependence patterns may ignore. In the previous sections, I hypothesized what some of these constraints might be and connected them to known patterns of causal selection judgments, arguing that they explain the patterns in question. In particular, I explored the patterns revealed by the experiments presented in Chapter 3. These experiments provide the first systematic study of causal selection judgments involving multivariate causes, and they revealed several surprising findings that are not easily reconciled with existing theories.

This raises the concern that the proposal made in this chapter might be too narrowly tailored to the judgments observed in the previous experiment—that it might be "overfitting" a very limited dataset. This worry is entirely valid and indeed unavoidable; it is quite possible that new experiments could demonstrate that some of the assumptions made here are untenable or fail to generalize beyond limited cases. For this reason, it is worth emphasizing that the project underlying the theory I presented is broader than merely explaining a small number of data points, and that it is not bound to the specific details of the present account. For each component of the theory, I have identified the main decision points and the choices that were made. These define multiple avenues along which the theory remains open to transformations. Amending the account along these avenues would preserve the fundamental insight that there is much to learn about causal cognition (and likely cognition more generally) by investigating the mental devices through which we concretely simulate occurrences of events.

Throughout this dissertation, I have also emphasized how this line of inquiry can be pursued without succumbing to the issues of indeterminacy typically associated with connectionist models. By grounding our hypotheses about internal processes in symbolic theories of mental representations, we ensure that our explanations remain tied to an identifiable structure, rather than getting lost in the opaque computations of purely network-based approaches. One might object that we can draw on such symbolic theories without committing to a full-fledged account of how they must be implemented in an actual cognitive mechanism. Instead, their insights could simply be integrated into classical descriptive models. Indeed, this is effectively what we did in Chapter 3, where we augmented an account based on Structural Causal Models (SCM) with a parameter w to capture people's propensity to reinterpret the rule in a manner consistent with plural negation. This parameter serves as a nod to the fact that people's causal reasoning depends on more complex internal procedures than a basic SCM can describe, yet it ultimately remains a descriptive shortcut: it flags a source of complexity without delving into the reasons for this complexity and the mechanisms by which it arises. In the end, however, a comprehensive account cannot remain agnostic about these matters. Merely

appending parameters to descriptive models does not explain why people adopt such interpretations or how these tendencies fit into the broader architecture of their causal knowledge. An approach that directly examines the cognitive machinery—the internal devices and processes through which people concretely simulate events—provides a way to address these questions head-on.

In the next part of this dissertation, I show how the approach developed here, which accounts for the patterns described in this chapter, naturally extends to a different (yet closely related) question: how do people use the explanations they generate as tools for learning from observations? In other words, having examined how individuals construct explanations from their internal representations of causality, I then look at how those same explanations guide and facilitate the acquisition of new causal knowledge.

Part II

Causal inference from explanations

Chapter 5

Introduction to Part II

In the first part of this dissertation, I proposed a theory of how people produce and evaluate causal explanations for an outcome. That theory addressed the question of how our explanations for an outcome E depend on the format and content of our causal knowledge. The present second part focuses on the related but converse question: how are the contents of our causal knowledge shaped by the explanations we are given for an outcome E? As in the first part, I concentrate on *causal selection* explanations—whereby one factor is singled out as particularly crucial for the outcome, without explicitly detailing the underlying mechanism.

Causal selection explanations are most deserving of study in some sense precisely because they do not detail any kind of mechanism. An explanation that details the mechanism, like "the car started because the key was turned, and turning the key turns on the engine" can itself be viewed as a "snippet" of a causal model, for example in the form of clause `Engine` $\leftarrow$ `Key`, so that the question of how such explanations contribute to learning amounts to the question of how a detached piece of knowledge can be overlaid onto one's pre-existing causal knowledge – for example, combining it with another snippet like `Car running` $\leftarrow$ `Engine` to form the more comprehensive model:

$$\{\texttt{Engine} \leftarrow \texttt{Key},\ \texttt{Car running} \leftarrow \texttt{Engine}\} \tag{5.1}$$

This is a relevant question in and of itself, but different from the one I want to look at in the following chapters. The role of causal selection explanations in learning is more mysterious because we have to understand how a mere *emphasis* on certain variables (by mentioning them preferentially over others), as in the mere statement that "the car started because the key was turned" can help a subject infer complex structures like (5.1).

This second part is divided into two main chapters, corresponding to two separate strands of work. In Chapter 6, I present joint work with N. Navarre, N. Bramley, and S. Mascarenhas, which was originally presented at the 2024 meeting of the Cognitive Science Society. The paper will appear here in full as it appeared in the proceedings. It briefly introduces some of the questions surrounding the role of explanations in inferences about causal structures. It proposes a novel experimental paradigm in which participants infer the causal rules underlying a certain set of data, where the data comes in the form of observation, and causal selection explanations about these observations. It asks participants to make inferences about causal rules in a more open-ended, higher-dimensional task than in previous studies on the subject. It provides evidence that explanations play a role in inferring causal rules even when the rules to be inferred are of some complexity, making the mystery outlined above more acute. It also presents a computational model of how explanations play into learning, based on the (pre-existing) idea that listeners infer causal knowledge from explanations by

reverse-engineering the causal theory behind an speaker's utterances, essentially asking: "what theory must the speaker hold about cars to favor `Key` as an explanation for the car's start?"

The second chapter is more strictly theoretical and constitutes new work written specifically for this dissertation. I take a step back from the assumptions made in the paper presented in the first section (as well as elsewhere in the literature), to re-examine the core issue from scratch: by what cognitive processes do explanations help us learn? I propose to criticize accounts based on reverse-engineering. In a nutshell, my argument is that reverse-engineering accounts fail to capture the notion that learning from explanations requires *less* work, not more work, than if explanations weren't available. Reverse-engineering accounts treat explanations as one more piece of data, on top of the observation that it explains. As such they prompt additional inference steps, on top of those prompted by the original observation, making the whole process computationally harder – in some versions, I'll argue, too hard to be realistic.

This critical discussion will lay the groundwork for a different proposal. I will argue that explanations simplify the process of learning from observations by focusing it onto relevant parts of the input, thereby reducing the search space for credit assignment. Trying to understand how a car starts by observing a driver start it is difficult, because so many variables may be potentially relevant. An explanation like "because the key was turned" helps because it tells the listener what to look at. This locates the lever through which explanations shape causal learning at a very low level, which can only be clearly articulated in terms of the connectionist architectures in which we have laid down our account in the first part. I will propose to model it in terms of an *attention mask* over the vector of inputs, and show how that model captures key patterns in the data collected from the experiment in the first section.

Chapter 6

Functional rule inference from causal selection explanations

Author note: This chapter is a verbatim reproduction of a paper published in the proceedings of the 2024 edition of the Cognitive Science Society conference, written in collaboration with Nicolas Navarre, Neil Bramley and Salvador Mascarenhas and with Nicolas Navarre and myself as first co-authors (Navarre et al. 2024).

6.1 Introduction

Humans form elaborate causal models of the world, which allow them to understand, forecast, and influence the events around them. A central puzzle in cognitive science concerns how these causal beliefs are learned. Extracting causal conclusions from observations of events is a notoriously hard problem (Bareinboim et al. 2022; Bramley, Lagnado, and Speekenbrink 2015), and everyday inference settings often provide few opportunities to perform the *interventions* (or experiments) needed to reliably infer the causal structure behind a distribution of events.

To mitigate these limitations, it seems plausible that people should frequently rely on social learning, to piggyback on the causal knowledge of one another in order to achieve an understanding of the world more efficiently. This lines up with the ubiquity of *causal explanations* in everyday discourse. From infancy, we frequently ask our peers for explanations for "why" things occur the way they do. As an everyday example, one might ask a neighbor 'Why did your flowers grow so well?' and learn something new from the explanation one receives (e.g. 'Because I used fertilizer').

A complicating feature of such everyday causal discourse is that explanations rarely lay out a complete mechanism sufficient to reproduce the explanandum, as we might expect from scientific textbook explanation. Rather, they tend to highlight one or a few of the causal factors involved and claim these as *the cause* of the event. You might point to fertilizer as the cause for the growth of these flowers, *rather than* for example the presence of the sun or water. This explanation seems reasonable in spite of the knowledge that sunlight and water are also prerequisites for flowers to grow, and would certainly have their place in an exhaustive causal theory of flower growth. Judgments of this kind, which single out a particular subset of causal variables as holding particular importance, are known in the psychological literature as *causal selection* (Quillien and Lucas 2023), or causal responsibility judgments (Lagnado, Gerstenberg, and Zultan 2013).

On the face of it, such selective explanations may appear to be poor conveyors of causal knowledge: by singling out a subset of variables in a system that often contains many more interrelated parts, they run the risk of reflecting only the explainer's preference for one kind of explanation. As such they might impoverish, rather than enrich, a requester's causal understanding. Yet, the psychological literature on causal selection has shown that people hold very consistent intuitions as to which of several events in a causal model is the most important cause of an outcome (Morris et al. 2018). This seems to unlock the possibility that people reverse engineer aspects of an explainer's beliefs about a causal system from the causal factors they choose to highlight in their explanations of specific outcomes.

We propose a novel experiment design to test this possibility. We put experimental subjects through a task of abductive causal inference, where they have to retrieve the causal structure underlying a dataset from a mixture of observational and explanatory evidence. Our design allows us to control the main known drivers of causal selection judgments. Our results show that causal-selection explanations help subjects generalize more adequately from limited data than when they are provided with observational data alone, or with other causal explanations that do not point at the main causes targeted by causal-selection judgments.

6.2 Theoretical Background

Recovering what structures and functional relationships underlie a system based on observations of the states of that system can be a particularly difficult task.

The crux of the challenge lies in the fact that all too often multiple distinct causal hypotheses might be equally compatible with a set of observations. Observing for example that an event E regularly follows the occurrence of two events A and B doesn't help me decide whether the underlying structure is one where the conjunction of A and B causes E to occur ($E \leftarrow A \wedge B$), of one where either one of A and B would have caused it to occur ($E \leftarrow A \vee B$) — here restricting the focus only to rules involving Boolean variables and connectives, for simplicity. A greater variety of observations might help narrow down the possibilities: if A occurs but B doesn't, while E still follows, I can exclude the possibility that A and B are both necessary for E to occur. I would still need additional observations however to rule out other possibilities, such as $E \leftarrow A \vee B$ or simply $E \leftarrow A$, or any other rule consistent with this limited set of observations.

The problem becomes all the more acute in settings where crucial observations occur infrequently. Whenever some of the events relevant to a causal system have very high or very low probabilities of occurrence, one rarely gets the chance to observe crucial event combinations. If A is always present, I cannot learn what the effect of its absence would be. Yet this is crucial information to infer A's causal relationship to E.

Remarkably enough, it is in those situations where the available observations are rather poor at covering the space of causal systems that causal-selection judgments are going to be particularly sharp, and exhibit strong preference in favor of certain variables.

6.2.1 Causal selection judgments

Causal selection judgments have been shown to be sensitive to two main factors: the causal *rule* that links candidate causes with the relevant outcome, on the one hand, and the *normality* associated with the different variables, on the other (Icard, Kominsky, and Knobe 2017; Morris et al. 2019; Quillien and Lucas 2023).

For example, in a situation where several different variables are each individually *necessary* for an outcome to occur (for example, when both water and fertilizer are required for my flowers to grow), subjects tend to think of the most unusual variables (the fertilizer) as 'the cause', and comparatively disregard the

importance of the most expected ones (the water), a pattern of judgment known as *abnormal inflation*. By contrast, in a situation where each variable is individually *sufficient*, people tend to favor the most probable events as explanations. This latter pattern of judgment is known as *abnormal deflation* (Icard, Kominsky, and Knobe 2017).

6.2.2 Counterfactual theories

Two successful theories of these patterns of causal judgment to date involve the notion of *counterfactual sampling* (Icard, Kominsky, and Knobe 2017; Quillien and Lucas 2023). According to counterfactual theories, causal-selection judgments involve a two-step process. First, one uses one's causal model to generate a sample of counterfactual situations, where the values of causal variables differ from what has actually occurred and the outcome potentially differs too. The frequency of each event across counterfactuals is a function of their prior probabilities, and whether or not the event effectively happened in the real world. The outcome of interest is determined by the events sampled, following subjects' causal models of the situation. Empirically, it seems that people consider counterfactuals that are both *likely* under the causal model of the situation, and *close to the observations they have made*. From a sample of counterfactuals one can compute a causal responsibility score, as some measure of the covariation between the states of causal variables and the outcome (different across models). The variables with the highest causal responsibility score are those that subjects are expected to favor in causal-selection judgments (Quillien 2020).

6.2.3 Inference from explanations

Not only do subjects show a lot of consistency in these patterns of judgments, they are also eager to assume that others follow similar patterns of judgments, and derive pragmatic inferences out of such assumptions. As shown by Kirfel, Icard, and Gerstenberg 2022, in a situation where two causes A and B are known to impact an event E, subjects told that E happened 'because of A' will infer that the underlying causal structure is $E \leftarrow A \wedge B$ when A is the more expected variable, and $E \leftarrow A \vee B$ when A is the more unexpected variable (in a situation where they are to choose between just those two structures). More broadly, it has also been shown that certain types of explanation serve as a guide to property generalization for both children and adults (Lombrozo and Gwynne 2014; Vasil, Ruggeri, and Lombrozo 2022).

6.3 Experiment

Here we present a novel experiment design, to show that this capacity to derive inferences from causal selection judgments can also help abductive causal inference in conditions closer to everyday life, where the causal structure to be guessed is of relative complexity, the space of possibilities open-ended, and the available observational data too limited to infer the rule with deductive certainty. This extends previous accounts of causal inference from explanation (Lampinen et al. 2021; Nam et al. 2023) by focusing on the role of causal-selection judgments specifically, rather than just considering the role of any causal explanations in inference.

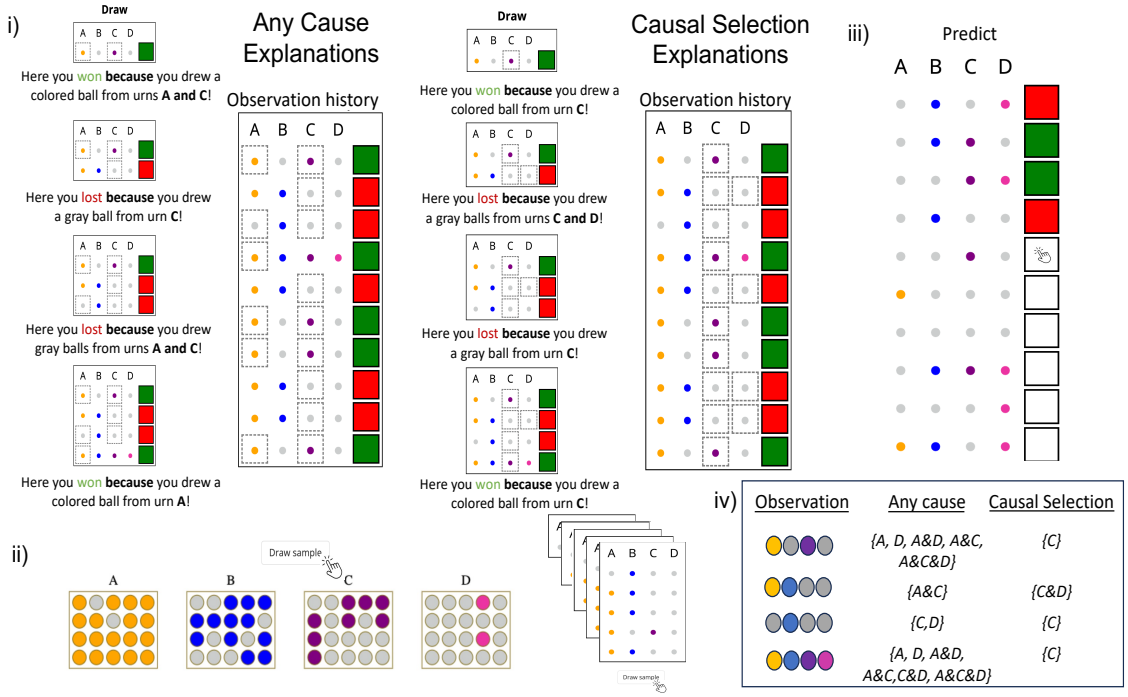


Figure 6.1: The experiment design; (i) shows a sequence of draws with all four samples in the experiment along with respective explanations. To the right of the samples is the entire history of observations once all 10 samples are drawn; (ii) shows the urns participants sample from and a visual example of the sampling process; (iii) shows the table of samples participants are tasked with predicting (not showing all 16). Participants have access to their observational history when making predictions. (iv) shows a comprehensive list of the observations and explanations given in the experiment.

6.3.1 Design

The game

Participants are invited to participate in a game where they must infer a hidden rule based on examples of winning and losing outcomes. The game involves four urns, each containing a mix of colored and uncolored balls, with each urn having a distinctive ball color, as in Figure 6.1-(ii). A round of the game involves drawing a ball at random from each of the four urns, with the result of these draws determining whether the player wins or loses the round. Whether a given draw from the urns corresponds to a win or a loss is determined by a fixed rule, but subjects are not told what the rule is. Their task in the experiment is to guess it.

The conditions

The experiment involved three between-participants conditions, two of which are depicted and described in Figure 6.1-(i).

In one condition, not depicted in Figure 6.1-(i) for reasons of space, participants only have access to observational data: they get to draw several times from the four urns and observe the outcome of each draw. To illustrate, on one particular draw, they might draw a colored ball from urns A, and C, a white ball from urns B and D, and then observe that this particular draw corresponds to a win, as in the first row of Figure 6.1-(i), minus the dotted squares. Such an observation gives them some information about the rule that links draws to outcomes. For example, it tells them that drawing a colored ball from urn D is not necessary for one to win in this game. We call this condition *observation only* or OBS.

In the other two conditions, participants see the same observations, but on top of that, they also have access to some *explanations* which tell them, for each draw that they observe, *why* they won or lost in that particular round of the game. These explanations point to a subset of the balls drawn as being *responsible* for the outcome of the game. These are the dotted squares around balls drawn from urns in Figure 6.1-(i).

In the *causal selection* condition, or CS, participants are given explanations that correspond to intuitive causal-selection judgments that a person knowing the causal rule would have been expected to make. We chose the relevant judgments for each observation by computing the predictions of two models known to provide a good approximation to human causal-selection judgments (see details below on causal-selection explanations).

In the *any cause* condition, or AC, participants are also given causal explanations for each observation, but these do not match intuitive causal-selection judgments. Instead, they point to any subset of the variables featured in an observation that played an active causal role, *except* for the one subset of variables which our best theories of causal-selection judgments predict to be the most important causes. The rationale behind this condition was to make sure that causal-selection judgments did not help subjects just by virtue of the fact that they point to any variable that made a contribution to the outcome, which could have provided a first step towards reconstructing the causal rule.

6.3.2 Materials

The observations that subjects saw consisted of random draws from the four urns represented in Figure 6.1-(ii). Probability acted as a proxy for normality in our design. Each urn contained a different mixture of colored and white balls, indicating the following probabilities of drawing colored balls from urns: $P(A) = 0.9$, $P(B) = 0.6$, $P(C) = 0.4$, $P(D) = 0.1$. The position of the urns was randomized across subjects, but for ease of exposition we will refer to those urns by the names in Figure 6.1-(ii). The use of urns allowed us to have a direct

handle on participants' subjective probabilities, a paradigm that has proved effective in past experiments on causal-selection judgments (Konuk et al. 2023b; Morris et al. 2018; Quillien and Lucas 2023).

The rule

The rule that determined the outcome, across all conditions, was as follows. To win, one must either draw a colored ball from both the high-probability urn A and the low-probability urn D, or from the intermediate probability urn C. In logical notation, this corresponds to

$$\text{WIN} \leftarrow (A \wedge D) \vee C. \tag{6.1}$$

We chose this rule because its logical form involves both conjunction and disjunction, so that we expect causal-selection judgments to be sensitive to both the abnormal inflation and deflation effects, as well as complex combinations of the two, depending on the target observation.

The Observations

In order to guess the rule in (6.1), subjects were provided with the 10 observations in Figure 6.1-(i). Participants drew observations successively from the urns by clicking the 'Draw sample' button above the urns. The order in which they appeared was randomized across subjects, but all participants saw functionally identical observations. Many of these observations were repeated, so that subjects only saw *four* unique observations in total, listed in Figure 6.1-(iv).

The choice of observations was constrained by four desiderata: (1) they had to be consistent with the probabilities implied by the urns, avoiding observations that the priors made too unlikely; the repetitions made sure that the frequency with which each color is drawn is proportional to its probability; (2) subjects had to draw a colored ball and a white ball from each urn at least once (to convey the idea that for each draw they observed, the alternative draw was a live possibility); (3) the two models of causal-selection judgments that we used as benchmarks had to agree as to the most important cause of the outcome for each observation (see next section for details); (4) for each observation, there had to be at least one active cause for the outcome (this mattered for the *any Cause* condition, see below).

Causal selection explanations.

The causal strength of explanations presented in the CS and AC conditions were computed using two models of causal-selection judgments, the Counterfactual Effect Size Model (CESM; Quillien and Lucas 2023) and the Necessity and Sufficiency Model (NSM; Icard, Kominsky, and Knobe 2017).

We considered these two theories because they have been shown to provide good predictions of subjects' judgments in a variety of documented cases (Quillien and Lucas 2023). However, our goal in this study was not to commit to one particular model of causal-selection judgments. Rather, we meant to probe whether explanations can help subjects in a causal-inference task without assuming a particular theory of how these explanations are generated.

In both theories, the causal responsibility of each event that influenced an outcome is a function of three main parameters: (1) the prior probabilities of drawing a colored ball from each urn, (2) the balls that were actually drawn in the case under consideration, (3) the causal rule that determines the outcome. They also include a sampling parameter s, which represents the extent to which the counterfactual worlds from which the causal impact of an event is computed are anchored to the actual world of reference. We reused the values

of that parameter that had been previously fit to behavioral data (Quillien and Lucas 2023), namely $s = 0.73$ for the CESM and $s = 0.15$ for the NSM.

Both models were run on each of the possible selections of variables. This includes the conjunctions of these individual candidate causes, (such as $A \wedge C$) or 'plural causes', as humans also hold consistent intuitions about such causal combinations, which are sensitive to the same factors as those driving judgments for singular events (Konuk et al. 2023b).

Given causal scores computed in this way, we selected the subset of draws whose causal score was highest as the explanation to be given to participants as explanations in the CS condition. Both models agreed on each of the four observations as to which event had the highest causal responsibility. The highest scoring events were highlighted with grey dotted boxes as in the Causal Selection Explanations table of Figure 6.1-(i). The explanations were also delivered linguistically to subjects for each observation as they drew them, as illustrated in the same figure.

Any cause explanations.

In the AC condition, the explanations we gave to subjects were sampled at random from any of the active causes of the outcome except for the one that was selected for causal selection judgments (see the full list in Figure 6.1-(iv)). Once an explanation had been provided for a given observation, we kept the same explanation for every repetition of that observation that a subject drew. The explanations were displayed in exactly the same way as in the CS condition. We took advantage of the fact that the locution 'X because Y' is one that can be used both for causal selection and for more generic causal attributions (Copley 2020).

Scoring inference

After subjects see the full ten observations, they are asked to make predictions for each of the 16 possible draws from these four urns. They are presented with a full table of possible draws as the one in Figure 6.1-(iii), with the boxes corresponding to the outcomes left blank. They should click on the boxes to turn them green or red, depending on what they think the outcome would be for each particular draw. We recorded the accuracy of each prediction made in this way, with subjects scoring 1 for a row if they gave a prediction matching what the outcome that the rule in (6.1) determines for that row, and 0 otherwise.

6.3.3 Procedure

We recruited 298 participants on Prolific from the United States, United Kingdom, and Canada. Each participant was randomly assigned to one of the 3 conditions (AC: 98, CS: 97, OBS: 103).

First we explained the mechanics of urn sampling and the relationship between the number of colored balls in an urn and the probability of drawing one. We had participants play a much simplified version of the game, involving just two urns, to illustrate how a rule mediated the relation between draws and outcomes, and the workings of the testing procedure that would follow. In the relevant conditions, we also gave them examples of explanations, making sure to pick examples were CS and AC couldn't differ, so as not to prime their interpretation of subsequent explanations one way or the other.

Participants were then invited to make ten "dry" draws from the urns, as in Figure 6.1-(ii) (right), where they weren't provided with any outcomes, so as to get them to internalize the probabilities associated to each urn. The draws were randomized so as to reflect the probabilities.

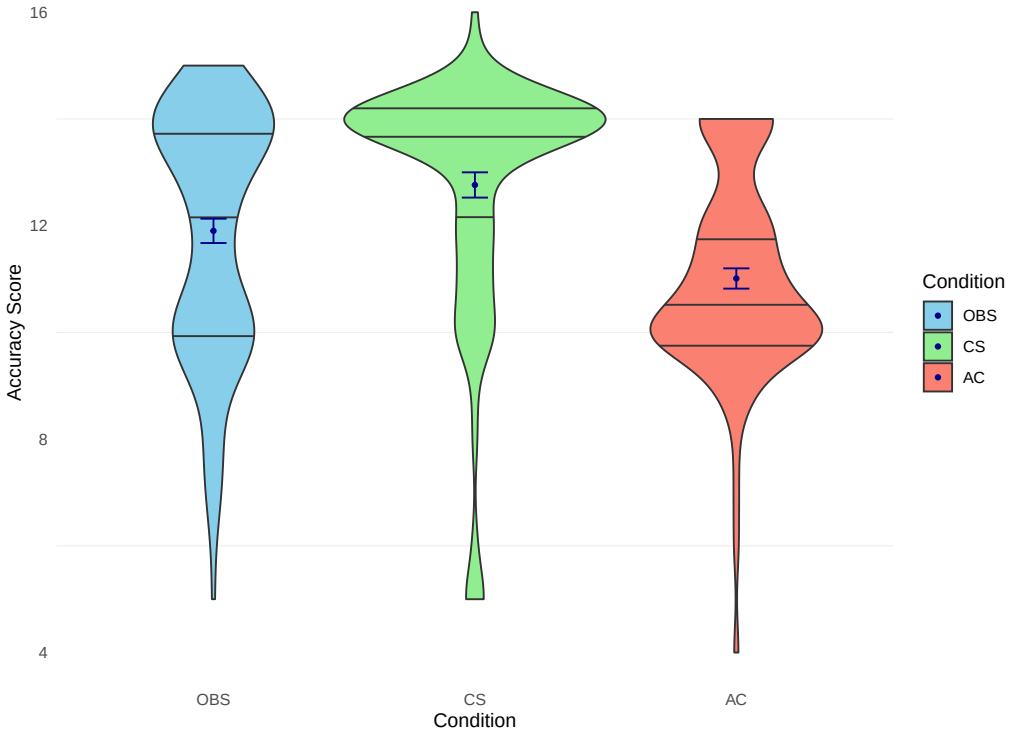


Figure 6.2: Prediction accuracy per condition. The blue dots represent the means per condition 12.76 (CS), 11.93 (OBS), 11.01 (AC). Error bars represent the standard error around the means. Solid black lines mark the medians and quartiles.

They then drew the ten observations with outcomes from Figure 6.1-(i), accompanied with explanations depending on the condition, and subsequently offered their predictions for all 16 possible samples. Finally, participants completed a brief questionnaire, where they had the opportunity to describe the rule they had in mind in prose if they so wished, and were asked some demographics questions, before being redirected to Prolific for payment. The experiment was programmed using the `JsPsych` JavaScript library (Leeuw, Gilbert, and Luchterhandt 2023).

6.4 Results

CS explanations helped subjects reach more accurate generalizations, while AC explanations made them less accurate

Participants' accuracy across all 16 samples is summarized in Figure 6.2. As suggested by the plot, subjects in the CS condition were overall more accurate than in either the OBS or AC conditions. A three-way ANOVA confirmed that the difference between the means of each condition was highly significant ($Df = 2$, $t = 15.65$, $d = 9.04$, $p < 0.0001$).

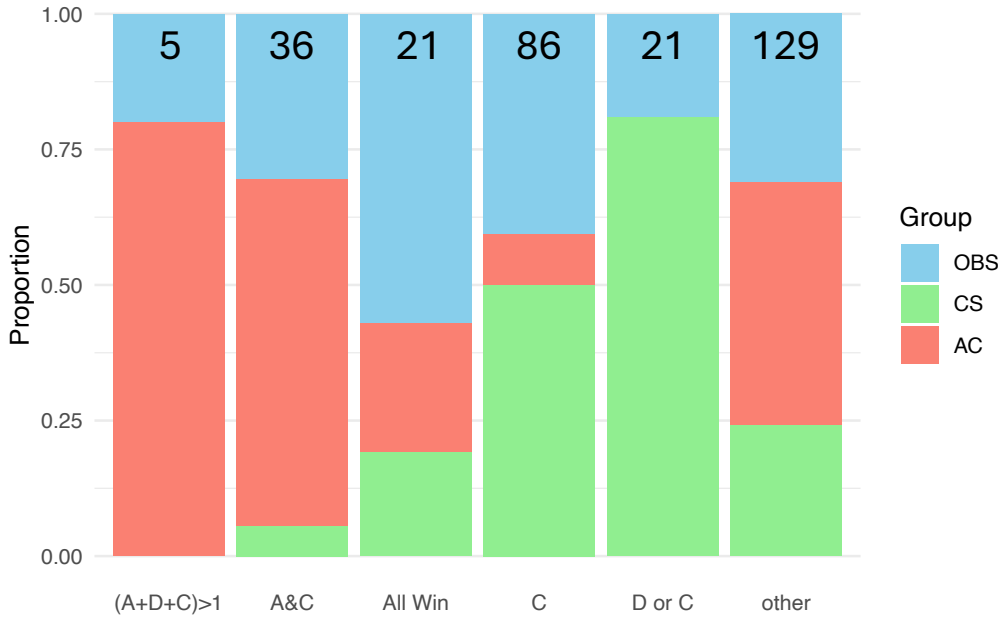


Figure 6.3: Most common rules inferred by the participants. The numbers in the bars represent the total number of participants in each group.

To further assess the effect of condition on subjects' accuracy, we ran a mixed-effects model using the accuracy of predictions for each row as dependent variable, the condition as fixed effect, and individual subjects as random effects. Results are summarized in Table 6.1. They confirmed the trend suggested by the figure: while CS taken as a factor had a positive effect on the correctness of guesses, AC had negative effect (compared to the OBS condition as baseline).

Removing the condition factor from the model resulted in a significantly worse fit to the data ($Df = 2$,$-\text{LogLik} = 2674.0$, $\text{BIC} = 5381.791$ for the full model; $Df = 4$, $-\text{LogLik} = 2690.1$, $\text{BIC} = 5397.057$ for the Intercept-only model; $\chi^2 = 32.205$; $p < 0.0001$), confirming that the condition subjects were placed in significantly affected their generalizations in the expected direction.

Subjects converged towards certain high-scoring rules

Because subjects gave an answer for all 16 observations possible with the available four urns, we were able to reconstruct the rule that guided their choices by looking at the truth table of their responses. Figure 6.3 plots the most popular patterns of responses per condition, translated into logical propositions that matched the contents of their responses. An outstanding pattern was the simple rule *C*, which was very popular in both the OBS and CS condition, although not in the AC condition.

Fixed Effects	**log-odds ratio**	**Standard Error**	**p-value**
Intercept (OBS)	1.11562	0.07304	$<$ 2e-16
AC	-0.29284	0.10178	0.00401
CS	0.31959	0.10707	0.00284
Random Effects	**Variance**	**Std. Deviation**	
Participant	0.1972	0.4441	

Table 6.1: Results of Mixed-Effects logistic regression: `Sample-Accuracy` $\sim$ `1 + Condition + (1| Participant)`

6.4.1 Computational Modelling

To help interpret the results from the experiment, we constructed a Bayesian model that makes inferences about possible rules from the observations and explanations. The model is comprised of two elements:

i) A prior probability distribution over all rules one could represent with four urns. We only retained deterministic rules consisting of Boolean combinations of colored and white balls from each of the four urns, giving us a total of 2^{16} distinct rules up to equivalence. A simplicity prior was applied to this probability distribution, which penalized rules whose definition depended on a greater number of different variables (Lu et al. 2008; Lucas et al. 2015).

ii) A likelihood function to update one's probability as a function of new observations and explanations. Observations simply update the probability by excluding rules incompatible with a given observation O and renormalizing probabilities over the remaining rules. For explanations, the model first computes a causal score for every possible explanation $E_m \in \mathbf{E} = \{E_1, \ldots, E_n\}$ and rule $R_m \in \mathbf{R} = \{R_1, \ldots, R_{2^{16}}\}$, by taking the square of the causal responsibility score $\kappa(E_m, O, R_m)$ that E_m would get for O, under the assumption that the correct rule is R_m, using the CESM model. Then, the model uses that causal score as the basis for the likelihood $P(E \mid O, R)$ of each explanation by normalizing over the causal score of all possible explanations, following equation (6.2).

$$P(E \mid O,R) = \kappa(E,O,R)^2 \left(\sum_{E_i \in \mathbf{E}} \kappa(E_i,O,R)^2 \right)^{-1} \tag{6.2}$$

The posterior distribution over rules is then updated based on how well these likelihoods predict the explanation that the learning model sees in each condition.

6.4.2 Model Results and Analysis

We compare model results in each condition on three dimensions: i) The Maximum A Posteriori hypothesis (MAP): which rule has the highest posterior after the four distinct observations; (ii) the position of the intended rule $(A \wedge D) \vee C$ in the posteriors' rankings; (iii) the weighted score of each distribution, i.e. the score that each rule in $\mathbf{R}$ gets in our experiments' test, weighted by their posterior probability.

	OBS	**AC**	**CS**
Obs. 1	323	15684 ± 182	22
Obs. 2	43	5985 ± 52	6
Obs. 3	37	2577 ± 23	4
Obs. 4	67	952 ± 10	14
MAP rule	C	—	C
MAP score	14.00	10.6 ± 0.037	14.00
Weighted score	10.00	10.36 ± 0.013	**11.65**

Table 6.2: Upper section: the posterior probability ranking of the intended rule $(A \wedge D) \vee C$ after each observation. Below: the rule that has the highest posterior probability (MAP) after the fourth observation in each condition, and the weighted score of the respective final probability distributions.

As shown in Table 6.2, CS explanations reliably rank the intended rule among the most probable candidate generalizations for the observed data, compared to the observation-only condition and the AC explanations on average. Additionally, the weighted score of the CS condition is significantly greater than that of both other conditions. Finally, it appears that the model accurately captures the attractiveness of the rule C, which stood out as the MAP in both the OBS and CS conditions, especially compared with the AC conditions, which favored a greater diversity of rules as MAP (with 5/60 AC configurations choosing C), in line with the distribution of that strategy across conditions, as reported above.

Another takeaway from the model results is that, even in the CS condition, the intended rule didn't come out as the MAP. Later iterations of this design will address this by examining cases in which we can expect CS explanations to guarantee the exact ground-truth rule underlying the game assuming optimal inferential abilities. In any event, these results concur with our experimental results in that the CS explanations bring the intended rule as the MAP much more reliably in light of new observations than the other two conditions.

6.5 Discussion and Conclusions

Our findings indicate that causal selection judgments serve as valuable cues when inferring causal structures from limited observational data. Our experiment is the first to provide evidence of this in a context where the causal rule underlying the data is of relative complexity and the space of possible hypotheses open ended.

Individuals not only demonstrate improved generalization from the data when causal-selection judgments are provided as explanations, they also exhibit notably poorer performance when presented with true but less selective causal explanations. These findings, in conjunction with the results from our computational model, strongly suggest that causal-selection judgments can aptly tap into humans' shared set of intuitions about causality to convey elaborate causal knowledge via relatively simple explanations.

Chapter 7

Inferences from explanation and attention

7.1 Introduction

The experimental paradigm presented in Chapter 6 revealed an important finding: that causal selection judgments can serve as valuable cues when inferring causal knowledge from observational data. This naturally raises a deeper question: *how* can an explanation that merely mentions a small subset of facts guide inferences about an underlying causal theory, whose content may be considerably more complex?

We proposed one answer, building on previous work (Kirfel, Icard, and Gerstenberg 2022), which suggests that causal selection judgments drive inferences by tapping into humans' shared set of intuitions about which factors are held more responsible, depending on the relevant causal rule of the domain. In this view, when a speaker singles out a normal or abnormal event as cause, they give cue as to which causal theory they have in mind, which a listener can interpret by considering what kind of theory would have led to single out this particular cause in the present context. While this account aptly captures many empirical patterns, and seems entirely realistic in the sort of simple situations that Kirfel, Icard, and Gerstenberg look at, scaling it up to contexts where the set of possible causal hypotheses is open-ended (as in our experiment, or *a fortiori* in many real-world contexts) is not straightforward.

In the following sections, I re-examine this question from the ground up and discuss in detail the sort of cognitive mechanisms that can in principle support these inferences. To set the stage, it is appropriate to begin with a concrete example first introduced by Kirfel, Icard, and Gerstenberg, which will help us capture the heart of the puzzle at hand. Suppose your flatmate Suzy just got the results of her last two med school exams, which she both passed. You know that she studied one of them (physiology) very well, but the other (anatomy) not as much, but you do not remember whether she needed to pass both exams or if just one was enough to get into med school. If Suzy then comments:

(1) "I got into med school because I passed anatomy".

You are drawn to infer that:

(2) Suzy needed to pass both exams to get into med school.

This inference "follows" in some way from the *abnormal inflation* effect, whereby abnormal causes are taken to be more important in conjunctive causal systems. Kirfel, Icard, and Gerstenberg provide some evidence to that effect, by showing that when the normal cause ("because I passed physiology") is cited instead, people tend to make the converse inference that Suzy needed to pass just one exam. But the exact logic of that *inference from explanation*, that is, the way in which (2) "follows" from the utterance in (1) via the pattern of abnormal inflation, is not at all self-evident. This is what we need to puzzle out. As a first approximation, we can describe such inferences from explanation in terms of a 'black box' function:

$$f_\xi : (\omega, \xi) \longrightarrow \mathscr{H}. \tag{7.1}$$

Where ω stands for an observation, here a pair (I, O) of input events (e.g. Suzy passed both exams) and outcome under focus (e.g. Suzy got accepted to med school), ξ for an explanation about ω, that consists of highlighting a subset of events in I, and $\mathscr{H}$ a space of hypotheses $\{h_1 \dots h_n\}$ itself consisting of causal functions $h : I \longrightarrow O$ that relate the input and output in ω. In the above example, the domain $\mathscr{H}$ could be instantiated as just two hypotheses:

$$\mathscr{H} \;=\; \{R_{\text{conj}}, R_{\text{disj}}\},$$

where

$$R_{\text{conj}} : M := A \wedge P \quad \text{and} \quad R_{\text{disj}} : M := A \vee P.$$

Where $A = 1$ corresponds to the proposition that Suzy passed the anatomy exam ($A = 0$ if she failed), and similarly P for "passing the physiology exam" and M for "getting accepted to med school". This high-level representation will serve as a roadmap for the discussion I will develop in the following paragraphs, where I try to progressively close in on the exact nature of f_ξ and the cognitive processes that realize it.

In **Section 7.2**, I start by formulating, in a semi-informal way, some conjectures about the information that f_ξ might engage as input, beyond the observation ω and explanation ξ. I distinguish three types of inferences, where each type engages a greater amount of information as input than the previous. To make inferences to the correct rule, we may indeed use:

1. Our own explanatory preferences, asking: "What causal theory in $\{R_{\text{conj}}, R_{\text{disj}}\}$ would have led me to say ξ?"
2. Take into account Suzy's explanatory preferences as well, asking: "What causal theory would have led me to say ξ, had I shared Suzy's normality assessments and explanatory style?"
3. Take into account, not only her preferences, but her communicative intentions. "What causal theory does Suzy want me to believe?"

This distinguishes three classes of cognitive processes which we can call, respectively, *reverse-engineering*, *mind-reading*, and *pragmatic reasoning*, each being a particularly complex special case of the previous. As I present these three classes, I will argue that inference from explanations reaches up to the second level of this hierarchy at best, and owes little if at all to the third. To clarify, my argument is not about the goals that motivate speakers' utterance, but about the information that is engaged in listener's inferences. I do not deny that communicative intentions are part of what drives speakers to produce explanations in the first place. I say that the cognitive process by which an explanation leads one to consider a certain causal theory as correct are not sensitive to considerations about the intentions of the speaker that delivered the explanation. This excludes the case where speaker's intentions give listeners reasons to refrain from using the explanation as input for inferences (if, for example, they are seen as deceitful). But insofar as listeners are making use of the explanation, they are not doing so in a speaker intention-sensitive way.

These distinctions will lay a useful groundwork for me to then in **Section 7.3** formulate more concrete hypotheses as to the nature of f_ξ, characterized at both a computational and algorithmic level of analysis (Marr 1982b): what are inferences from explanations trying to achieve and by what processes do they achieve it? Their goal can and should be characterized more precisely than "getting at the right causal theory", in particular in contexts like that of our experiment when reasoners do not have enough information to make sure they arrived at the right theory with certainty. In these cases we may want to ask what is a second best option. An answer to that question partly determines our notions of how the inputs (ω, ξ) and hypotheses $\mathscr{H}$ are represented in our minds, and what processes relate them to draw inferences f_ξ.

The first hypothesis I will consider is that inputs and hypotheses are represented in a symbolic format as full-fledged causal rules and that the function that relates them is an instance of Bayesian inference. Its goal is to bring listeners probability distribution over possible rules closest to that of speakers. I will argue that this is not a realistic possibility – even assuming strategies to approximate these inferences – beyond a few simple special cases. This is because the reliability of such inferences depends on a reasoner's ability to track the set of alternative causal hypotheses and alternative causal explanations for each observation, both of which scale exponentially with the number of dimensions in the data.

This will pave the way for me to develop an alternative view, that sees the inference as carried out at a subsymbolic level. My view is that inference from explanation is not conducted separately from the more basic inference

$$f_\omega : \omega \longrightarrow \mathscr{H}. \tag{7.2}$$

from the observation ω to causal hypotheses that might capture it, but happens against the backdrop of inductive error-driven mechanisms by which f_ω is carried. I will propose to understand explanations as *instructions* as to *which part of the input I* in ω should be attended to as one engages in inductive learning from ω. That is, the inference f_ξ should be seen simply as some refinement of f_ω where the input I in ω to be attended to is narrowed down and focused onto the most relevant variables. By doing so, explanations alleviate the frame problem burdening this kind of inductive learning, that is the problem of deciding what part of the input is relevant to f_ω. This proposal captures a strong pre-theoretical intuition that isn't captured by reverse-engineering accounts, which is the intuition that explanations make it *easier*, not *harder* to make sense of observations.

More specifically, I propose to understand f_ω as familiar process of gradient backpropagation, in a feedforward network where the activations of every neuron depends on the activation of the inputs neurons via activation functions dependent on weight parameters Γ. The input-output pair ω is used to induce the correct function $h : I \longrightarrow O$ by a gradient-descent process that moves every parameter γ up or down in the direction that minimizes the distance between the activation value of neurons representing O and the value that O takes in the observation ω, when inputs are themselves initialized to the value they have in ω. The shape of this process means that any change in input activations should influence the assignment of credit to parameters down the line. Explanations Ξ exploit this lever, by inflating the activation values of every variable associated with neurons mentioned in each $\xi \in \Xi$, and deflating (or nullifying) that of other neurons. This process nudges a learner to converge to functions $h : I \longrightarrow O$ that give a greater importance to the variables ξ. I present a model that operationalizes this principle in terms of an *attention mask* α applied to the input vector of the networks as it engages in gradient-based learning from ω. I show how such a model can capture key patterns in the data presented in the experiment of the previous chapter with very minimal assumptions about the underlying network's architecture.

Readers can see how the shape of the present account of inferences from explanations parallels our account of the production of explanations presented in Chapter 4. Just like we proposed that a subject's most preferred explanation involve variables that make a larger contribution to the outcome in that subject's internal

model, here we propose that explanations lead the same variables to make a larger contribution in the internal model of the listeners that decides to make use of them in learning. I will also show that this parallel allows for a different use of explanations in learning, closer to existing reverse-engineering accounts – although I will not develop this lead in as much detail here. The way in which we determined the causal importance score $\kappa(C,O)$ implicit to a network N in Chapter 4 makes $\kappa(C,O)$ end-to-end differentiable with respects to the weight parameters Γ of N. This means that for every pair (ω,ξ), every weight parameter γ of the network can be moved up or down in the direction that increases $\kappa_{\omega,\xi}(C,O)$ where C is the set of variables mentioned in ξ and O the outcome observed in ω. I suggest that this could implement the sort of reverse-engineering processes supposed by Bayesian accounts, but without falling prey to issues related to the complexity of Bayesian inference.

7.2 Three Levels of Inferences from Explanation

Level 1: Reverse-engineering

One way to understand the inference that leads us from Suzy's explanation in (1) to the conclusion that she needs to pass *both* exams to get into med school is to assume that people internally reverse-engineer Suzy's causal knowledge, by considering what sort of causal theories might have led Suzy to favor that explanation. This reasoning strategy could *in principle* be applied at scale, for any similar inference from explanations. This would involve internally running a version of Algorithm 1 below. The description in

Input: ω: an observation, ξ: an explanation about ω.
Initialize $\mathscr{H}$, a set of candidate causal hypotheses compatible with ω;
foreach $h_i \in \mathscr{H}$ **do**
  Build an internal simulation model that assumes the ground truth causal rule is h_i ;
  Task this model to produce an explanation for the observation ω;
  Record the explanation generated as ξ_i;
end
return the set $\{h_i \in \mathscr{H} : \xi_i = \xi\}$;

Algorithm 1: Reverse-Engineering Explanations

Algorithm 1 remains voluntary very permissive about the sort of simulation model that is involved, or about the way in which it may produce explanations. In that it highlights some very general conditions that must hold for reverse-engineering causal theories from explanations to be a tractable process.

First, the hypothesis space $\mathscr{H}$ must be sufficiently small to be iterated over. It is easy to see why this implies that $\mathscr{H}$ cannot contain *all* causal hypotheses that are in principle compatible with ω in the general case. In practice, a reasoner will have to rely on heuristics to narrow down the set $\mathscr{H}$ of all compatible hypothesis to a smaller set $\mathscr{H}'$ that is still representative enough of $\mathscr{H}$ (see e.g., Bramley et al. 2016, for a presentation of strategies for doing so). The moment one restricts $\mathscr{H}$, however, there arises a second potential problem, which is that there may be no h_i in the subset $\mathscr{H}'$ such that $\xi_i = \xi$. A natural solution might be to move from a purely eliminative filter (step (1) in Algorithm 1) to a *probabilistic* approach. In other words, replace step (1) by:

(1') Pick whichever $h_i \in \mathscr{H}$ maximizes $P(\xi \mid h_i)$,

or even better by:

(1") Update your probability distribution over $\mathscr{H}$ via Bayes' rule using $P(\xi \mid h_i)$ as the likelihood term.

I will return in Section 7.3 to a more detailed discussion of this probabilistic perspective. For the moment, let us note one more important fact about the feasibility of Algorithm 1, this time negative: it does *not* crucially depend on the fact that the explanation ξ is produced in a *communicative* context. This is made clear when one considers that, for example, I could have found the explanation "Suzy got into med school because she passed anatomy" as a note written in the diary of Mary, our other flatmate. I could have still used that explanation as a means to make inferences as to whether Mary needed to pass both or just one exam to get into med school. And doing so would have involving executing the same steps (1)–(3) in just the same way. The only difference is that I may have doubts as to whether Mary knows the conditions of acceptance to med school as well as Suzy does. But such doubts would only lead me to adjust my *trust in the results* of the process in Algorithm 1 – whether I can take the output it delivers as the right causal theory. It would not however change the results themselves, or the process that led to them. In other words, the communicative intention behind ξ may be entirely orthogonal to how we run the *reverse-engineering* procedure in Algorithm 1. This is a distinction that it is important to emphasize. The fact that speakers often mean to communicate some causal knowledge through their explanations, does not automatically mean that the explicit consideration of such intents plays a role as such in the inferences that listeners draw from those explanations.

Level 2: Mind Reading

Now, the identity of the speaker may still, in principle, affect how we use their explanations in inference. This is most evident in situations involving false belief. Suppose, for example, that I know (but didn't tell Suzy) that the anatomy exam at her med school is graded very leniently, whereas the physiology exam is graded very severely. Thus, although Suzy believes that she is more likely to pass physiology than anatomy, in fact I know the opposite to be true. Given how causal selection explanations depend on normality, I might want to take into account this discrepancy as I reverse engineer Suzy's causal knowledge from her explanation.

This example illustrates a second layer of inference from explanations: I might reverse-engineer causal rules not solely based on my own normality assessments and explanatory tendencies, but also based on those of my interlocutors. To do so, I can use my knowledge about which events people ordinarily consider as normal or abnormal, and the sort of conventions that they ordinarily follow as they explain things. An iterative algorithm can then leverage that knowledge at scale by incorporating it into a variant of Algorithm 1 that tunes the reverse engineering process to an interlocutor's priors and explanatory preferences, as presented in Algorithm 2 below.

Input: ω: an observation, ξ: an explanation about ω,
Π: parameters encoding the interlocutor's priors and explanatory preferences.
Initialize $\mathscr{H}$ a set of candidate hypotheses for the correct causal theory;
foreach $h_i \in \mathscr{H}$ **do**
  Build an internal simulation model that assumes the ground truth causal rule is h_i ;
  Set the default parameters of the model that encode normality priors and explanatory tendencies to Π ;
  Task this model to generate an explanation for the observation ω;
  Record the explanation generated as ξ_i;
end
return the set $\{h_i \in \mathscr{H} : \xi_i = \xi\}$

Algorithm 2: Perspective-Taking Reverse-Engineering

Similar remarks as for algorithm 1 apply regarding the possibility and relevance of a probabilistic variant of this algorithm. Regardless of the probabilistic or not nature of the algorithm, a question it raises is: how much of such perspective-taking are people really capable of? A host of research in cognitive and social science indeed suggests that reasoning on the basis of another subject's assumptions comes with a lot of difficulty, and that humans tend to engage in it only reluctantly, and often insufficiently (e.g. Birch and Bloom 2007; Epley et al. 2004; Horton and Keysar 1996). Often, a subject will implicitly reason about others' beliefs as if under the assumption that they are the same as one's own, even when they can explicitly recognize that assumption to be inaccurate. In the case of causal explanation, this would amount to rely on their own priors and explanatory tendencies (i.e. engaging in a process like that of Algorithm 1) despite knowing that this approach is sub-optimal. Indeed, results from Kirfel, Icard, and Gerstenberg indicate that people are engaging in such simplifications. Their subjects were asked both to produce explanations assuming a certain causal rule (e.g. Suzy needs to pass both exams) and to infer rules when given a certain explanation. Those whose judgments diverge from the populational patterns of abnormal inflation and deflation tend to produce inferences that match less robustly with the expected pattern. That is, if an individual tends to cite the normal cause ("Physiology") as an explanation assuming Suzy needed to pass both exams, they are less likely to infer that she needed to pass both exams when told that she got into med school because of Anatomy. This does not mean that people were unaware of the different between conventional patterns of explanations and their own – as evidenced by the fact that groups of subjects with unconventional explanation styles still veered towards the expected inference on average. But it suggests that people were reticent to use their knowledge about conventional patterns of explanation even when they were in a position to do so. In terms of the levels distinguished in this section, subjects in Kirfel, Icard, and Gerstenberg's experiment engage in a mixture of Level 1 (simple reverse-engineering) and Level 2 (mind-reading) reasoning.

This is striking because it means that even when the conditions are ideally suited for Level 2 reasoning, as in the Suzy vignette – where there are only two candidate explanations and causal hypotheses to track, so that people could in principle explicitly spell out the explanation Suzy is expected to give assuming each candidate causal theory – people still feel the need to engage in some amount of Level 1 reasoning as a shortcut. This, in turn, suggests that as the complexity of the inference increases (because the ground-truth rule is more complex, and/or there are more causal variables at play), reliance on Level 2 reasoning should decrease even further.

Finally, it is relevant to note that the second level of reasoning described here, still does *not* meaningfully depend on the communicative nature of the explanation. That is, whom Suzy addressed her explanation to, or whether Suzy she intended it to be addressed at all does not affect the kind of inference I make regarding the underlying causal structure. I could, here again, be reading the explanation in Suzy's secret diary and the inference I draw via either of Algorithms 2 would remain unchanged. The fact that subjects are reluctant to engage Level 2 inferences even when they do not depend on a consideration of speaker's intentions argues against the idea that they might want to engage in inferences that further depend on a consideration of such intentions on top of that. I look at such "Level 3" inferences below.

Level 3: Pragmatic Reasoning

Suzy's communicative intentions might in principle become relevant to my inference if they can serve as cues to her underlying causal knowledge. One might reason as follows: "Suzy said ξ, and she has reasons to believe that uttering ξ is the best way to lead me to infer that h_i is true. Hence, she intends for me to adopt h_i, implying that this is the model she herself considers correct." A version of the process that might underlie such *pragmatic inference* is represented in Algorithm 3, inspired by rational models of pragmatics (Frank and

Goodman 2012; Goodman and Frank 2016). Note that Algorithm 3 is a simplified sketch of what it would take to perform inferences based on a speaker's communicative intentions. A fully detailed model would require comparing the utility $U(\xi^{h_i})$ of the explanation ξ for Suzy under the assumption she believes h_i to its utility assuming other alternative hypotheses, and also to the utilities of all alternative explanations ξ' that Suzy might have produced but did not. This would allow one to convert utilities to probabilities in a more principled manner.

Input: ω: an observation, ξ: an explanation about ω, Π_1: parameters encoding Suzy's priors and explanatory preferences; Π_2: some model of Suzy's beliefs about my own interpretative tendencies.
Initialize $\mathscr{H}$;
foreach $h_i \in \mathscr{H}$ **do**
  Construct an internal model **of Suzy**, denoted Suzy^i. It has all the attributes captured in Π_1 and Π_2, complemented with the supposition that Suzy believes h_i to be the true causal rule. This may be relaxed to a probability distribution $P^i(\mathscr{H})$ over candidate hypotheses.;
  Compute the probability $P(h_i \mid \xi)$ that **I the speaker** would assign to h_i as I hear ξ, according to Suzy. This amounts to running Algorithm 2 with the normality priors and preferences **that Suzy assumes I possess**, as encoded in Π_2;
  Evaluate a utility $U(\xi^{h_i})$, defined as a function of the discrepancy between the probability distribution induced by ξ over $\mathscr{H}$ and Suzy^i's own distribution over $\mathscr{H}$.
end
Convert the ranking of $U(\xi^{h_i})$ across hypotheses in $\mathscr{H}$ into a probability distribution over $\mathscr{H}$, assuming that a higher utility indicates a greater likelihood that Suzy intends h_i to be inferred;
return the updated probability distribution over $\mathscr{H}$;

Algorithm 3: Pragmatic Reverse-Engineering

Even absent this additional layer of complexity, such pragmatic inferences should be challenging. The fact that we are able to take into consideration the communicative intentions of speakers for our interpretation of some statements of ordinary discourse does not automatically imply that we can do the same for the interpretation of any statement. It all depends on the complexity of the processes needed to take communicative intentions into consideration. In this particular case, one can see why doing, even in the simplified manner proposed in Algorithm 3, will be very difficult, because Algorithm 3 includes Algorithm 2 as a subcomponent.

Additionally, reverse-engineering Suzy's causal knowledge based on how useful she might find a certain explanation for the goal of conveying causal knowledge to me assumes that these explanations are effective at conveying causal knowledge to begin with – or communicative utility would not have been a factor for why she chose a certain explanation rather than another. But this itself becomes very unclear, the moment we recall that the same communicative goal could be met by simply stating one's causal theory explicitly, as in:

(3) "I needed to pass both exams to get into med school."

In other words, with respects to their respective utility for the goal of conveying causal knowledge, causal selection explanation are usually in competition not just with one another but also with other explanations of the more full-fledged kind exemplified in (3). The fact that the causal rule is often more complex and harder to state than (3) cannot be a counter-argument here, given that the complexity of the pragmatic inference would itself also grow as the rule becomes more complex. In fact, we have given many reasons to think that it would grow even faster that the costs of detailing one's causal theory upfront.

Hence we cannot make causal selection judgments a cost-effective strategy for communicating causal theories without assigning a very high cost to the length of utterances while assuming no additional cost for the extra processing efforts required to make sense of more laconic explanations. In other words we would have to assume that thinking is always much cheaper than talking. Yet not only is this assumption counterintuitive, but the very existence of social learning (of which learning from explanations is a special case) is in some sense predicated on the fact that the exact opposite is true.

In the next section, I delve deeper into the complexities that come with the attempts to implement the kind of Bayesian inference process implied by Algorithm 3 or by probabilistic versions of the simpler Algorithms 1–2. This paves the way to an alternative view, according to which the point of explanations is precisely to reduce the processing efforts involved in ordinary inferences from observations.

7.3 The Nature of the Inference

In the previous section, I argued that people's inferences from explanations do not involve pragmatic reasoning stricto sensu but may instead rely simpler reverse-engineering processes. I now examine the hypothesis that the reverse-engineering processes outlined in Algorithms 1–3 are instances of Bayesian inference. This idea, which aligns with earlier work by Kirfel, Icard, and Gerstenberg (2022) and with the proposal presented in Chapter 6, suggests that inferences from explanation are about using reverse-engineering processes like those described in the previous section to associate a likelihood term to a focal explanation ξ given each of several causal hypotheses, and then use that term to update one's posterior expectations over those causal hypotheses. In this section I give a brief argument that performing such inferences at scale—beyond relatively simple contexts like those in Kirfel, Icard, and Gerstenberg's experiments—is very challenging. I then propose two alternatives that rely on the connectionist representations of causal models discussed in Part I of this dissertation.

7.3.1 Inference from explanation as Bayesian Inference

The idea that inferences from explanations were a special case of Bayesian inference entails that then the posterior probability $P(h \mid \omega, \xi)$ of any hypothesis $h \in \mathscr{H}$ given an observation ω and an explanation ξ can be decomposed following Bayes' rule:

$$P(h \mid \omega, \xi) = \frac{P(\xi \mid h, \omega)\, P(h)}{\sum_{h' \in \mathscr{H}} P(\xi \mid h', \omega)\, P(h')}, \tag{7.3}$$

where $P(h)$ is the prior probability of hypothesis h, and $P(\xi \mid h, \omega)$ is the likelihood of producing explanation ξ given that h is true and that ω was observed. This decomposition seems to fit nicely with the reverse-engineering process outlined in Algorithms 1–3. These algorithms simulate and record, conditional on each hypothesis in $\mathscr{H}$, an explanation or a causal importance score. So it seems that they can also generate a the likelihood term $P(\xi \mid h, \omega)$ on the basis of which posteriors can be computed via (7.3). Looking more closely, however, this presents a major difficulty.

In general, we know what explanation we would give, or how attractive we find a certain explanation, conditional on a certain causal theory. But that does not mean we know which explanations we are *likely to give* and how likely we are to give them. Accordingly, none of the existing theories of causal selection generate an actual *probability* for explanations given a causal theory and normality assessments. Instead, they yield qualitative *preferences* over certain explanations in the form of causal importance scores $\kappa(\xi, \omega)$

assigned to each explanation given a certain observation. For $P(\xi \mid h, \omega)$ to have any meaning, it must be interpreted as "the probability that a subject will produce ξ as an explanation for ω, given that they take h to be the true causal rule." Yet the causal importance score $\kappa(\xi, \omega \mid h)$ is not equivalent to that propensity, and cannot trivially be interpreted as such. This is most evident in the fact that importance scores do not naturally sum to one over the space of possible explanations.

One might argue that these scores could be converted into probabilities via normalization (or a softmax transformation)—an assumption we adopted in the model presented in Chapter 6 . But this strategy raises a considerable issue of scalability, if it is to be taken as a theory of how these likelihood terms are computed concretely. To normalize the causal importance scores $\kappa(\xi, \omega \mid h)$ across all $\xi \in \Xi$ into a probability distribution $P(\Xi)$, one would have to compute a score for *every* explanation that a speaker could, in principle, produce rather than just the one uttered. This is particularly problematic in view of the fact, argued at length in Part I, that plural causes are fully fledged candidate explanations, whose combined importance cannot be trivially reconstructed from a simple linear combination of parts. This means that the number the number of candidate explanations $\xi \in \Xi$ is on the order of the powerset $\mathscr{P}(\mathrm{I})$, where I is the set of distinguished variables in the input data I for the pair $(I, O) = \omega$.

One alternative might be to engage in repeated iterations of Algorithms 1–2 for each hypothesis h_i, recording each time the explanations $\{\xi_i^1 \ldots \xi_i^n\}$ that come to mind, and finally counting the proportion of ξ_i among those such that $\xi_i = \xi$ to estimate the probability $\mathscr{P}(\xi \mid h_i)$ indirectly. Yet this strategy merely trades off complexity for precision, resulting in a likelihood term with substantially lower evidential quality. Moreover, empirical studies on causal selection judgments systematically reveal high variability, which would compound with the noisiness inherent to such a process to further undermining the reliability of such internally sampled likelihood estimates. Because of these challenges, a tempting alternative is: why not use the causal importance scores—or the subjective preferences for various explanations given a causal rule—*directly*? In other words, instead of trying to retrieve a likelihood term $P(\xi \mid h_i, \omega)$ for each $h_i \in \mathscr{H}$ in Algorithms 1–3 above, simply record the causal importance score $\kappa(\omega, \xi \mid h_i)$ directly, then after all $h \in \mathscr{H}$ have been iterated over, skipping any normalization steps simply **return** the subset of $h \in \mathscr{H}$ with the highest $\kappa(\omega, \xi \mid h_i)$.

This alternative is appealing because it circumvents the need to normalize over an enormous space of potential explanations. Note however that by embracing this option we depart from the realm of Bayesian inference. An inference that attributes evidential weight to quantities that are not probabilities is by definition not Bayesian and the corresponding update process won't be decomposable via Bayes' rule. Below I propose one way to implement it in terms of gradient-based learning procedures over neural networks. It exploits certain properties of the $\kappa(C, O)$ measure defined in Chapter 4 of this dissertation, in particular the fact that it makes $\kappa(C, O)$ end-to-end differentiable with respects to the weight parameters of a neural network. This is not the main account in terms of attention masks announced in the Introduction as the main focus of this chapter. I develop that account a bit later in Section 7.4.

7.3.2 Explanations as Nudges to Inductive Inference

I have presented arguments against viewing Bayesian inference as a plausible mechanism for using explanations to infer causal structure, at least at scale. It is possible that fixes or approximations can be proposed that overcome the challenges described above. However, in what follows, I will set aside the effort to "rescue" Bayesian inference, and instead offer an alternative view of how explanations play a role in causal learning.

A common theme to the accounts proposed so far is that they ground the inference from explanations to causal hypotheses in a separate inference from hypotheses to explanations. This makes inferences from explanations a special case of "inference to the best explanation," with the explanations themselves treated as

explananda and the causal structures (plus observations) as the *explanans*. This reversibility is natural in the context of Bayesian inference, where each conditional probability $P(h \mid e)$ can be expressed naturally in terms of the converse $P(e \mid h)$. The account presented in this section in this sections takes a different approach and assumes that these inferences are carried out via a different inference procedure, such as a gradient-descent based learning procedure over a neural network, like the one described in Algorithm 4 for inferences based on observations alone. From this standpoint learning causal theories from examples amounts to repeated applications of Algorithm 4 with different observations, gradually adjusting the network parameters so that the mapping from input nodes to output nodes mirrors the relevant causal relationships.

Input:

- $\omega = (I, O)$: an observation that pairs input events I with outcome facts O.
- N_0: a feedforward neural network whose input layer V_{in} comprises one node per variable in I, and whose output layer V_{out} comprises one node per variable in O.
- Γ_0: initial network parameters (either randomly set or derived from prior learning).

1: **Initialize** the network N_0 with parameters Γ_0, **then do:**
2: **Step 1: Input Activation** Set the activation values of the input nodes to reflect I (e.g., assign $+1$ for an event that occurred, -1 for an event that did not occur).
3: **Step 2: Target Specification** Set the target activations of the output nodes according to the outcome variables in O.
4: **Step 3: Forward Pass** Propagate activations through the network to obtain predictions for the output layer.
5: **Step 4: Loss Computation** Compute a loss function (e.g., mean-squared error) by comparing the predicted output activations to the target activations.
6: **Step 5: Backward Pass** Perform backpropagation to update all network parameters $\gamma \in \Gamma$ in the direction that minimizes the loss function.
7: **Repeat** until some convergence criterion is met or a maximum number of iterations is reached.

Algorithm 4: Gradient-Descent Learning from Observations

In moving from an initial N_0 to some idealized "fully converged" N_n, a learner passes through a series of intermediate network parameterizations $\{\mathsf{N}_1, \ldots, \mathsf{N}_n\}$, each representing a partially learned approximation of the ground-truth causal function. Each N_i on that trajectory can be viewed as its own function h_i mapping input activations to output activations. Consequently, a set causal selection explanations $\xi_1^{\omega}, \xi_2^{\omega}, \ldots, \xi_n^{\omega}$ for any given observation ω can be generated from each of those networks. This is true regardless of which theory of causal selection we adopt, but it is worth noting that the mechanisms for generating these judgments that we proposed in Chapter 4 makes it easier to generate explanations out of those intermediate networks and use them to help the learning process, in two ways.

First, it does not require generating counterfactuals from each intermediate network N_i to arrive at an explanation. Every intermediate network along the learning trajectory *de facto* comes with an imbalance set of weights that highlight which variables currently exert the strongest effect. Hence people might directly compute relevance scores over each intermediate scores, which spares them the need to explore neighboring counterfactual states to "stress test" the system before they produce an explanation. Note also that it makes less sense to stress test an internal representation of a system if we have reasons to think that our representation is still very imperfect, as when we are still learning about it. In that context, computing the explanations

implicit in our representation serves a different purpose, which is to align those explanations with the ones we are given by some external source.

Second, the assignment $R_i(C, O)$ of relevance to a cause C for outcome O is a function of the continuous parameters Γ_i of each N_i in such a way that it is differentiable with respect to each parameter $\gamma_j \in \Gamma_i$ in any given network state. Since also $\kappa_i(C, O)$ merely divides $R_i(C, O)$ by $\mathscr{C}(C, O)$ which is fixed for any network in a given state. This means that we can, for all intermediate networks $\{\mathsf{N}_1, \ldots, \mathsf{N}_n\}$ and observations ω compute partial derivatives

$$\frac{\partial \kappa_i(C, O)}{\partial \gamma_j},$$

and thus track down how increasing or decreasing a parameter γ_j would alter the causal importance assigned to C, by altering the flow of relevance to variables in C. Taken together, these two features allow us to incorporate explanations directly into the learning pipeline of Algorithm 4, by enriching the loss function to include a term that nudges each parameter in the direction that would increase the flow of relevance to inputs in C. Concretely, we can define an enriched version of Algorithm 4, presented below as Algorithm 5.

In this enriched training procedure, the network learns not only to predict the correct outputs O, but also to respect the constraint that the variables mentioned in ξ carry a certain level of causal importance. The differentiable nature of $\kappa_i(C, O)$ ensures that standard gradient-based learning techniques can integrate these explanatory constraints seamlessly into the optimization loop. Repeated application of Steps 2–7 gradually molds the network's internal parameters so that it satisfies both *predictive* and *explanatory* objectives in tandem. This type of learning pipeline has precedents in the machine-learning literature, being used to learn from labeled datasets (e.g. Ross, Hughes, and Doshi-Velez 2017; Zaidan, Eisner, and Piatko 2007), albeit using different criteria and metrics for explanations.

Such an approach can be seen as an implementation of the strategy outlined in Section 7.3.1: it uses causal importance scores *directly*, as nudges that guide a reverse-engineering process towards models that assign greater importance to variables mentioned in ξ. It does not require turning that these scores be first normalized into probabilities. A complete theory of how we use explanations to learn causal theories would likely include such a reverse-engineering mechanism. Yet I will not explore this strategy any further here, a task I will leave to future work. Instead, in the next section I want to focus on another way in which explanations might guide inference, which I take to be of greater theoretical import. It departs altogether from the reverse-engineering accounts outlined in the previous subsections, to propose that explanations' role in causal is to guide inferences from observations by instructed the learner what features of the data to focus on.

7.4 An Attention-based account of Inference from Explanation

A feature common to the accounts outlined so far is that they see learning from explanations as an extension of the methods by which we learn from observations, applied to a different type of data. Because explanations are not *facts* but rather *discourse about* facts, this requires a mechanism to extract from explanations some information in a format such that it can be handled by the same processes by which we learn from observed facts. The reverse-engineering algorithms described above include a way to extract an evidential term (be it in the form of a conditional probability or of a loss gradient) for each ξ, so that the same inferential machinery (Bayesian inference or gradient descent) that we use to learn from observation can use ξ to update our causal knowledge. It is not clear however why we should want learning from explanations to be handled by the same sort of mechanisms by which we learn from observations. Such accounts seem to rest on the unspoken premise that learning from observations is an easy or straightforward enough process to begin with, so that

extending the same mechanisms to explanations "come for free" in a theoretical sense.

But this premise is flawed, in that learning from observations alone is itself a very difficult process. And attempting to fold explanations into that same inference pipeline via reverse-engineering only compounds the difficulty. Hence the alternative proposal I want to offer here: that explanations help causal learning, not by offering more material for inference, but by making inferences from the same material (that is, observations) *easier.* To place this proposal in context involves recalling a central source of difficulty in inference, which is the one that I will argue explanations help alleviate. This source of difficulty is often discussed under the heading of the "frame problem" (Fodor 1983; McCarthy and Hayes 1969). It can be summarized in the following way: in an environment with many variables at play, only a small subset of those variables are *truly relevant* to an inference; yet, there is no universal method to identify that subset *before* performing the inference.

To illustrate, suppose you are watching tennis matches and seek to understand what it takes to succeed at tennis, starting with only a novice's understanding of the sport. You observe a wealth of data in the form of points, sets, and entire matches. Observable inputs include the players' actions and dispositions, the conditions of the court, and so on; the outputs are who wins and who loses. These conditions already put you in a great spot for learning, as you have enough data to distinguish statistically significant patterns. You may however easily be overwhelmed by the sheer number of variables that could potentially affect each outcome—did the player score a point due to the speed of their serve, their footwork, their mental composure, or something else entirely? If we take learning as a credit assignment procedure to variables based off some metric (like a loss gradient), we have to decide what are the variables whose loss gradient we keep track of. Tracking all "potentially relevant" factors is not an option, because the relevance of factor can only be assessed if we track it in the first place.

Part of the point of causal models is precisely to alleviate this problem, by specifying which variables are conditionally independent of others. (Pearl 2000; Spirtes, Glymour, and Scheines 2000). If a variable β (say, the brand of a player's headband) is known to have no relation to an outcome, except through a mediator variable μ (headbands only help by blocking sweat, hence they only matter when weather is very hot) or some confounder χ (better players are sponsored by fancier brands, but the brand does not add any value by itself) the learner can ignore β to focus on μ or χ as they assign credit for the outcome. Yet even with such partial help the problem does not disappear. In practice, humans often restrict their inferences to *submodels* of the system at hand (Fernbach, Darlow, and Sloman 2011; Fernbach and Rehder 2013), suggesting that even a known causal structure may be selectively pruned when its scope is large. Moreover, the advantage offered by conditional independences is less available in a *learning* context, where precisely a stable causal model is yet to be built.

Now suppose you came to the tennis court with your cousin Bill, who is a tennis expert and offers real-time commentary: "He won this point because of his serve" or "She lost that rally because she lost her temper." Each utterance effectively directs your attention to a small subset of features, and conveys the presumption that those features are particularly relevant for explaining the outcome. This process can be incorporated straightforwardly into a pipeline like Algorithm 4 by means of an *attention vector* that selectively amplifies or attenuates certain input activations.[1] This results in the enriched learning algorithm described in Algorithm 6. It does not require any additional loss term (unlike in Algorithm 5); instead, each explanation ξ directs how input activations are weighted during training, thus biasing the learner to find a solution that leverages those

[1] Here, "attention vector" refers to a simple masking or weighting scheme on the input nodes. This differs from the more sophisticated self-attention and multi-head attention architectures used in modern sequence models (Bahdanau, Cho, and Bengio 2014; Vaswani et al. 2017). Our usage treats attention as an explicit cue or mask that is externally provided by the explainer, rather than learned end-to-end by the network.

features. Such guided learning procedure does not introduce a separate explanatory inference step, because the explanation modifies *how* the original observation ω is processed, rather than requiring an additional inference from ξ. In doing so, it reduces the learner's search problem by telling them which components of I are most relevant, thereby mitigating the frame problem in a top-down manner. Before we get into more modeling details, to show how such a process captures the effects of explanation we observed in our experiment, it is useful to see at a high level why we should expect explanations to benefit causal learning in this framework.

By increasing the activation values associated with the variables $\beta \in \xi$,we amplify the gradient updates on the parameters that depend on those variables. These stronger updates will push the learner's model in a direction that brings it closer to the explainer's model. Assuming the explainer knows best, this dynamic should help us learn a good model faster and more reliably. For instance, suppose I naively noticed over the last few games that the louder a player screamed during a serve, the more points they won. With no prior tennis knowledge, I might hypothesize that *screaming* is a key factor. Imagine now that cousin Bill is here to tell me that "they won because of their strong serves". This would redirect my attention to serve speed, a more relevant feature which just happened to be correlated with screams. Once I have correctly identified serve speed as a key factor and learn to weight it accordingly in my model, I'll be able to naturally de-emphasize superficial correlates like screaming. Indeed much less credit will flow to that variable as I repeatedly observe that all its co-variation with wins can be explained away in terms of serve speed.

In a different vein, consider a subsequent twist: I watch a strong server go on a losing streak. Without further guidance, I might infer that perhaps I had overrated serve speed and downgrade its importance drastically. If however this observation comes with Bill's comment that "He lost because he's bad on clay," I can credit these losses to a different variable (surface type), which mitigates the extent to which I would backtrack on my theory about serve speed. Here again, assuming Bill's domain knowledge[2] is sound, the explanation saves me from the mistake of interpreting as negative evidence observations that are imputable to other factors, which I might have otherwise overlooked.

7.4.1 Model Implementation and Application to Data

I now want to introduce a model of causal learning designed to operationalize precisely the attention-based account of explanation learning outlined above. Here I focus on high-level features of the model. Readers can find detailed code implementation in the public repository at https://osf.io/zv2m5/. Specifically, the model will be applied to the same task participants faced in the experimental paradigm described earlier. In addition to the theoretical considerations regarding *implementability* and *scalability* developed in the previous section, my goal here will be to show that the attention-based account *also* provides a better account of empirical data on people's inference behavior under complex rules. In particular, the account can explain two key patterns in our dataset that the reverse-engineering approach proposed in Chapter 6 struggles to accommodate. One is the observed popularity of the hypothesis that the rule of the game is `W:= C`. The other is the fact that people's performance decreased in the ANY CAUSE condition – where people where given explanations that

[2]Interestingly, note that explanations coming from speakers with only *partial* knowledge of the domain might be helpful as well. Suppose for example that Bill the tennis expert goes to the court with Suzy, the medical student. Although she may not have as much general tennis knowledge, her knowledge of Anatomy and Physiology make it easier for her to catch on the fact that a player's recently injured shoulder prevents them from hitting strong backhands, and offer that fact as explanation for his loss of certain points. This can in turn lead Bill to attend to a variable that he may have otherwise missed, wrongly attributing these losses to some other factor.

A complete account of such integration of insights from local domains of expertise into broader theories by means of explanations is probably very difficult to give. Nonetheless the present account seems better positioned than accounts of the symbolic-bayesian kind that see inference from explanation as updating posterior probabilities over rules represented as wholesale causal theories.

did not correspond to causal selection judgments – as compared to the OBSERVATIONS ONLY condition.

The Popularity of the Rule `W:= C`. Our experiment revealed that many participants found the rule

$$\texttt{W := C},$$

which says that a player wins a round if and only if they pick a colored ball from urn `C`, a very attractive as an explanation for the data, not only in the OBSERVATIONS ONLY condition but even more so in the CAUSAL SELECTION condition where those observations where accompanied with *causal selection* explanations – see Figure 7.1 below, reproducing the same figure presented in the previous chapter. Intuitively, this is not surprising given that *three* of the four unique explanations in that condition explicitly mentioned *C* (and no other variable) as the factor causing the outcome. Figure 7.2 recalls the explanations for each condition.

Yet this intuition is difficult to reconcile with a purely reverse-engineering approach, when one also considers that for the *fourth* observation, participants were given as explanation

(4) "You lost because you drew a colored ball from urns `C` *and* `D`."

This makes it odd for nearly half of the participants to still endorse `W:= C` alone as the game's rule if they were inferring rules by directly reverse-engineering the given explanations. Consider by analogy a situation where are four switches `A`, `B`, `C`, `D` in a room but I know that *only* switch `C` controls the light. In that context, there would be no situation in which I would say:

(5) "The light is off *because* switches `C` and `D` are off."

From which it follows in a reverse-engineering account of explanations that if I am a houseguest at some house and hear my host utter (5), I know that the probability

$$P((5) \mid \texttt{Light:= C})$$

that he would have said (5), were the rule `Light := C` accurate, is equal to zero. Hence by Bayesian inference the posterior probability $P(\texttt{Light:= C} \mid (5))$ would be driven to zero as well, or to some very low probability if one add an assumption that there is always some residual probability to any utterance. And the exact same reasoning should apply to the rule `W:= C` in the context of our experiment after people have heard the explanation in (4). In the model presented in Chapter 6, the fact that a reverse-engineering model could still predict that `W:= C` should be an attractive rule in the CAUSAL SELECTION condition crucially depends on the following two additional modeling assumptions.

1. *A simplicity prior:* An inductive bias favoring simpler rules biases inference toward `W:= C` over more complex rules that combine `C` with other variables. This is a reasonable assumption, and one that is supported by independent evidence (e.g., Bruner, Goodnow, and Austin 1956; Feldman 2000), but it would not by itself be sufficient to compensate for the extremely small likelihood of $P(\texttt{W:= C} \mid (4))$

2. A *revisionist* version of the counterfactual theories (CESM and NSM) we used to generate the likelihood function.

   These theories indeed restrict the set of candidate causes whose counterfactual dependence profile will be computed is restricted to the variables that classify as causes in a categorical sense to begin with. If `W:= C` is the ground truth rule, then `D` is not a causal factor at all, so the cause $\texttt{C} \wedge \texttt{D}$ that underlies the explanation in (4) should not even be considered to begin with. Instead we made the assumption

that people as they reverse-engineer causes consider the causal importance score of every combination of variables, even those that would not even categorize as an actual cause in a categorical sense. by considered a causal But this is a highly non-standard use of the models.

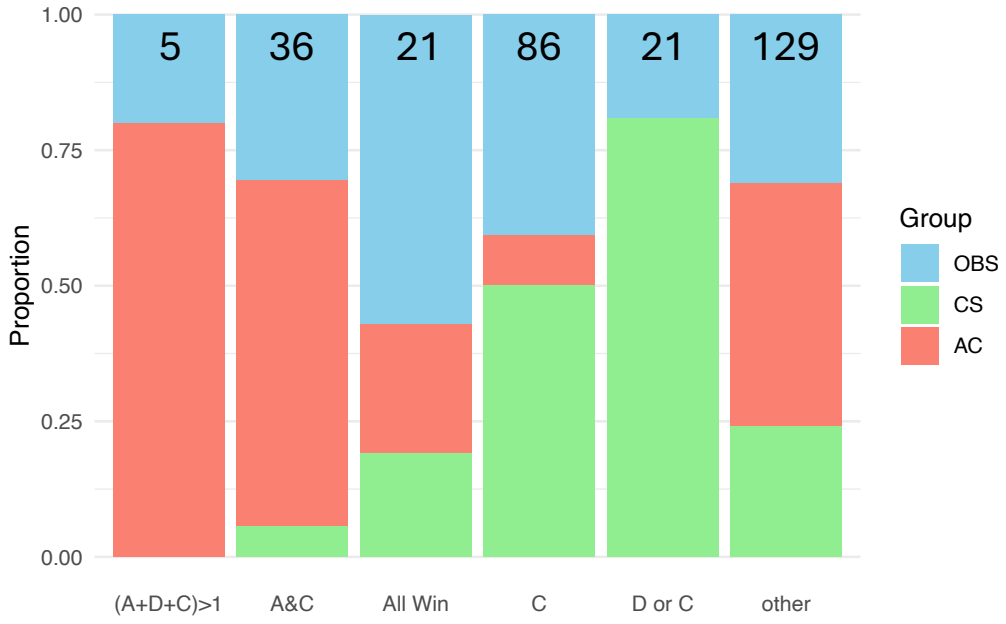


Figure 7.1: Most common rules inferred by the participants. The numbers in the bars represent the total number of participants in each group.

Predictive Performance in the ANY CAUSE Condition. Even setting aside the theoretical issues pointed out with the above assumptions, by which the model can reconcile the attractiveness of the rule `W:= C` with a Bayesian model, it is also worth remarking that they force the model into another corner. Under these assumptions, the model predicts a better performance in the ANY CAUSE condition than in the OBSERVATIONS ONLY condition. This goes directly against our results, which show that participants in the ANY CAUSE performed worse than even the participants that weren't given any explanation at all. It is surprising that a model based on reverse-engineering would do that. If explanations are used as pointers towards the rules that are most likely to have generated them via causal selection processes, we would expect explanations that do not match causal selection judgments to always be misleading. It is only because the model has relaxed the likelihood function associated with explanations so much, via the assumption 2 above, and by softmaxing over causal scores in a lenient enough way that high-scoring rules incompatible with the given alternative explanation do not get excluded.

An alternative account based on attention. I show in the following that an account of causal selection explanations as instructions to direct the focus of the learner is well equipped to capture these same patterns,

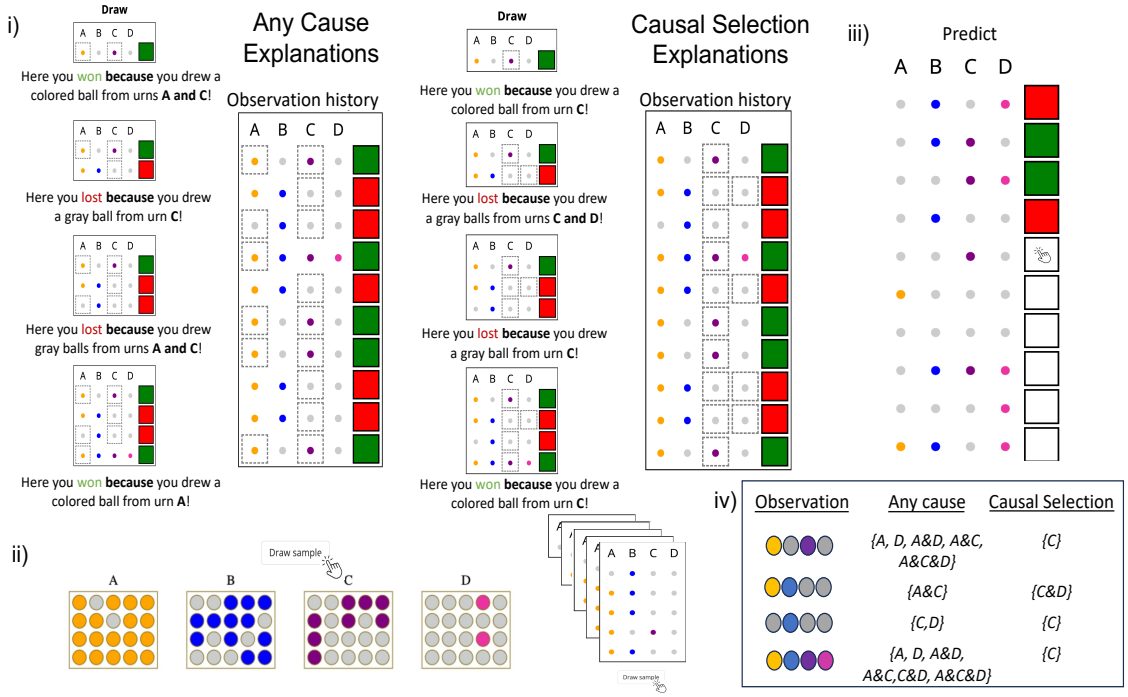


Figure 7.2: The experiment design; (i) shows a sequence of draws with all four samples in the experiment along with respective explanations. To the right of the samples is the entire history of observations once all 10 samples are drawn; (ii) shows the urns participants sample from and a visual example of the sampling process; (iii) shows the table of samples participants are tasked with predicting (not showing all 16). Participants have access to their observational history when making predictions. (iv) shows a comprehensive list of the observations and explanations given in the experiment for each condition.

and more generally the results of our experiment. To provide some high-level intuitions first, it makes sense why many participants in the CAUSAL SELECTION condition still find the rule `W:= C` attractive after hearing (4) as an explanation. It is because (4) simply *highlights* variables `C` and `D` for that particular trial. This gives the learning process a nudge in the direction of rules that give more importance to `D` as it sees this observation. But the incremental nature of the process, which adjusts internal parameters rather than removing certain hypotheses from consideration altogether or giving them a low probability, still leaves open the possibility that the process converges towards a model that makes the outcome depend on `D` alone, i.e. equivalent to the rule `W:= C`, without requiring any additional assumptions (not even a simplicity prior). And this possibility is likely to be further encouraged by the emphasis put on `C` alone in the other three explanations. Because such an account based on attention does not need to make any special assumptions to capture the attractiveness of `W:= C`, it will also be in a position to capture the fact that people are worse off by receiving the explanations they are given in the ANY CAUSE condition.

Incidentally, note that the attention account does not predict that explanations different from standard causal selection judgments should *always* hinder learning performance. It allows for other kinds of explanations to still be useful as long as they highlight sufficiently important factors. This flexibility is another advantage of an attention-based view, and it sheds light on how explanations from a speaker with only *partial* domain expertise (as is the usual case) can nonetheless be useful.

The attention mechanism I propose below is embedded in a neural learning framework. Given that neural models can come in many different formulations, I should first specify the assumptions I made about the network architecture and learning process. In formulating the design of the model, I adhered to a general principle of *minimalism*: when presented with equally valid options, I select the simpler, less parameterized alternative—even when it may be less psychologically realistic. This is because my main objective here is to provide a proof of concept. I want to isolate and highlight the impact of the attention mechanism itself. The simplifications I make ensure that the observed learning trends of the model are attributable to the attention mechanism itself rather than to extraneous hyperparameters. I show that even a minimalist implementation captures the key trends in our data that the reverse-engineering model fails to explain based on attention alone. I also flag potential extensions that could increase psychological realism in future versions.

7.4.2 Basic learning pipeline

Network architecture. The task involves observing multiple samples from four urns. Each sample is a draw of exactly one ball per urn, combined with an outcome (win or lose). Because the direction of causality is clear (from urn draws as inputs to game outcome as output), and because inputs from different urns are assumed independent, I assume a **feedforward neural network** N with no recurrent or lateral connections. The input layer V_{in} of the network comprises four neurons A, B, C, D, each corresponding to one of the urns that subjects drew from in the experiment. The output layer V_{out} contains a single node W corresponding to the game outcome (win or lose). Figure 7.3 schematically illustrates this architecture, including a single hidden layer (discussed below). Because each variable (white/colored ball, win/lose outcome) is binary, I let the neurons in hidden and output layers have **activation values in** $[-1,1]$, and the hyperbolic tangent $\tanh(\cdot)$ as activation function. The default activation value for input neurons is in $\{-1,1\}$ (+1 for a colored ball, -1 for a white ball) and can later be modulated by attention (potentially stretching outside of the $[-1,1]$ boundaries), as I will explain below.

The choice of $\tanh(\cdot)$ and activation values in $[-1,1]$ is not itself dictated by the structure of the task design, but it does not fundamentally alter learning capacity compared to more conventional sigmoid or ReLU networks and has the advantage of aligning with the convention used in the CILP framework (see, e.g.,

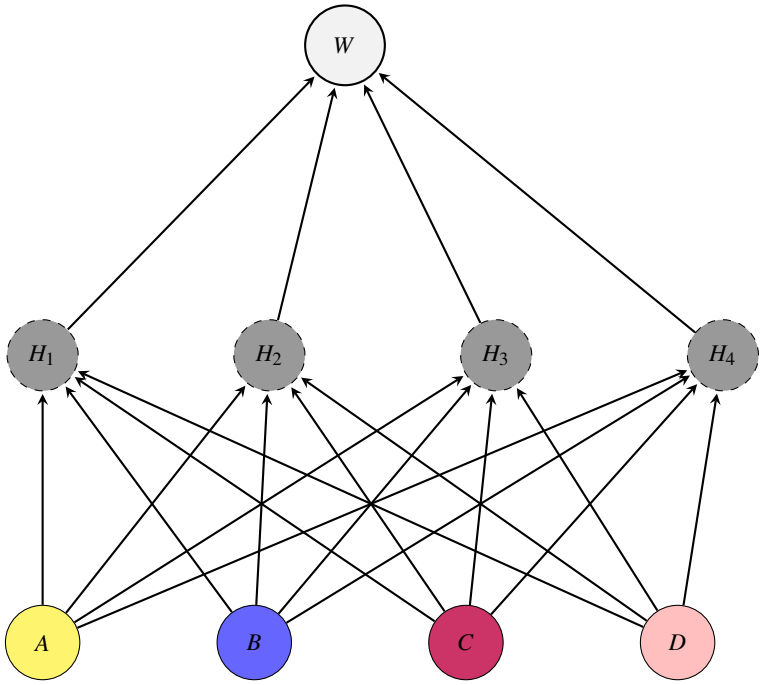


Figure 7.3: A schematic feedforward neural network with four input nodes (A, B, C, D), four hidden neurons (H_1, H_2, H_3, H_4), and a single output node W. All connections are directed from bottom (inputs) to top (output).

Garcez, Lamb, and Gabbay 2009) which we have used in Chapter 4, which helps with interpretability. It also make it simpler later on to implement attention mechanisms uniformly, in terms of a simple product applied to input nodes' activations (which would have been harder to do if negations had zero-valued activations).

I also posit a **single hidden layer** between input and output nodes – which is sufficient in principle to learn any boolean function over the space of inputs with enough data (Cybenko 1989). In the spirit of minimalism announced in the introduction, I assume that the hidden layer has the same dimensionality as the input data itself — i.e. that it contains **four hidden neurons** in total. A more psychologically realistic may have assumed fewer hidden neurons than there are input neurons in the data. Fewer hidden neurons impose a stronger "information bottleneck" (Tishby, Pereira, and Bialek 1999), effectively forcing the network to learn compressed or more abstract representations of the input, which fits with evidence that human learners favor simpler rules when explaining data (see, e.g., Bruner, Goodnow, and Austin 1956; Feldman 2000). For present purposes, I chose to run the model with a four-neuron hidden layer structure, to make sure that the results obtained owe nothing to this kind of implicit simplicity bias.[3]

Loss function and parameter updates At each observation, the model computes the difference between the network's output (the activated value of the output neuron) and the target ($+1$ for a win, -1 for a loss).

[3] I experimented with versions of the model with fewer hidden neurons, and the overall trends in the data reported in this section were not affected by such design choices.

We define the loss L as the squared error in a single observation ω:

$$L_\omega = \left(y_\omega - \hat{y}_\omega\right)^2,$$

where y_ω is the target (± 1) and $\hat{y}_\omega$ is the network's output for the observation ω. The error is not averaged across a batch of observations but computed for each observation in succession, in the random order with which they present themselves to the subject. Each subject in our training model is trained, not just on the four *unique* observations but on the entire 10 observations effectively presented to each subject, including repetitions.

The absence of batches might not be entirely psychologically realistic, because it seems plausible that, although the order of presentation of observations has an effect, people also keep an eye on several observations at a time as they try to extract a pattern – especially so in a design like ours, where the record of all past observations was kept on the screen for subjects to see. We omit considerations of batch level tracking however, again in an effort to simplify the model as much as possible.

The gradient of L with respect to each parameter w_i (weight or bias) is computed via standard back-propagation (Rumelhart, Hinton, and Williams 1986), yielding a *loss gradient* $\frac{\partial L_\omega}{\partial w_i}$ that indicates how each parameter should change to reduce the error. Parameters are then updated via the simple gradient descent rule:

$$w_i \leftarrow w_i - \frac{\partial L_\omega}{\partial w_i}.$$

I altogether omit regularization terms or momentum or even a learning rate. Again, while such hyperparameters are standard in machine learning, I prefer to exclude them here to keep the model as minimal as possible.

7.4.3 Simulated subjects and training epochs

Having outlined the basic neural learning pipeline, our next step is to simulate multiple "subjects," each corresponding to a single run of the training procedure. The model then generates predictions on the test set of 16 observations that was presented to subjects (all possible combinations of the four binary variables), and we measure its predictive accuracy to compare it to the accuracy observed for subjects in a certain condition in our data.

Training duration. Our empirical data indicate that participants—despite seeing the same stimuli—vary widely in their final performance and in how long they take to respond. A natural interpretation is that while all individuals rely on the same underlying learning process, they differ in how *long* they run that process (i.e., how many epochs of parameter updates). Some participants invest more time or cognitive resources into trying to come up with an inference. To capture this variability, we sample the number of epochs E for each simulated subject from an exponential distribution with rate λ:

$$x \sim \text{Exponential}(\lambda), \quad E = \max(1, \text{round}(x)).$$

The exponential distribution with rate λ has and expected mean

$$\mathbb{E}[X] = \frac{1}{\lambda},$$

and a cumulative distribution function

$$F(x) = 1 - e^{-\lambda x}, \text{ for } x \geq 0.$$

This sampling process is equivalent to sampling the number of epochs from a poisson distribution with $\frac{1}{\lambda}$. Intuitively, $\frac{1}{\lambda}$ reflects the average "extra" time or effort a participant invests. Larger values of λ indicate that participants saturate more quickly (fewer epochs), while smaller values of λ imply that some individuals train for many epochs. A central feature of the exponential distribution is its *memoryless* property. Formally, if a random variable T is exponentially distributed, then for any $s,t \geq 0$,

$$P(T > s+t \mid T > s) = P(T > t).$$

In other words, the conditional probability of "still running" for an additional time t does not depend on how much time s has already elapsed. This distinguishing feature of the exponential distribution explains why it typically allows much larger deviations from the mean than, for instance, the binomial distribution with similar means. Such behavior makes sense when we think of participants as successively weighing conflicting intuitions and observations against a burgeoning hypothesis. If a participant's observations quickly reinforce a single, consistent causal story, the learning process "congeals" and the participant stops pondering further. However, each new observation has some probability of contradicting the prior story, which triggers further reevaluation. This cycle continues as long as contradictions appear, thereby extending the reasoning process—analogous to waiting times in a memoryless queue.[4]

Empirical motivation from RT Data. The assumption of an exponential distribution of epochs is also independently empirically motivated by the data on response times we collected in our experiment. Figure 7.4 illustrates the distribution of response times in our experiment for the relevant trials. We fitted and compared various distributions (normal, exponential, ex-Gaussian) using `gamlss`, as shown in the code accompanying this dissertation. As is commonly found in cognitive tasks (e.g., Heathcote, Popiel, and Mewhort 1991; Ratcliff 1978), the response times follow an ex-Gaussian pattern. This is often interpreted as a sum of (i) a Gaussian component for basic perception/action latencies, and (ii) an exponential component for the variable thinking time. By matching our epoch sampling to an exponential distribution, we reflect the idea that individuals differ in how extensively they process the same data.

Simulation Pipeline. Algorithm 7 summarizes how we simulate multiple subjects, each with a randomly sampled number of training epochs. After training, each model predicts outcomes on the 16 test configurations, and we record the accuracy (the fraction of correct predictions out of 16).

In this pipeline, the rate parameter λ is fitted against the data, representing the only free hyperparameter in addition to the central *attention* parameter α described in the next section. The key idea is that, although all subjects share the same core learning process, inter-individual differences in invested effort or time emerge as different numbers of training epochs. Although introducing an exponential rate λ formally adds a hyperparameter, it also broadens the scope of our model (when compared to the reverse-engineering model presented in Chapter 6 for example): rather than explaining only the *mean* performance in each condition, we

[4] Another, more organic way to generate exponential waiting times for our training process would have been to design a conditional stopping rule, where the process has no preprogrammed end but keeps feeding itself observations from the training data until it reaches a high enough level of "confidence" by some metric. This would make the model analogous to a Drift Diffusion model, but with the drift rate being a product of the internal learning dynamics rather than a mere model parameter. Later versions of this model should probably involve such a mechanism. For now, I simply bake in the exponential waiting time via a rate parameter for convenience.

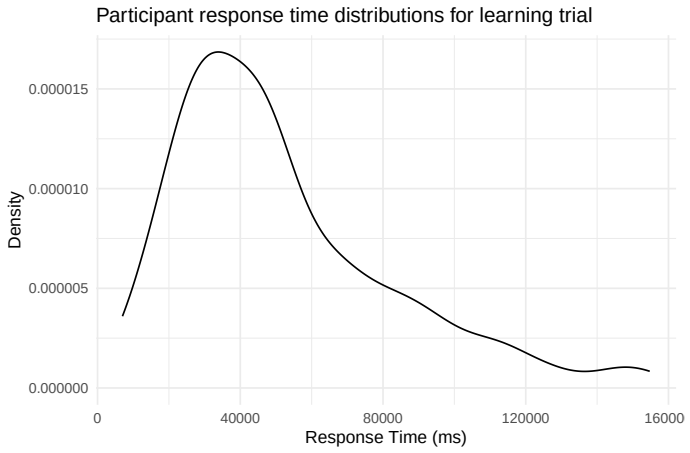


Figure 7.4: Distribution of time spent by subject on the experiment trial where they were presented with the observation relevant for inference, showing the characteristic long tail. Similar distributions are observed for other crucial trials, such as the one where people give their predictions. Model fit metrics confirmed that the ex-gaussian distribution provided a better fit to the distribution than Gaussian or Poisson alternatives. Note also that I excluded from this distribution one participant who took over 30 minutes (< 2000000 ms), under the supposition that they might have been a "true" outlier. Reincluding them would have made the long-tailed pattern even sharper.

aim to account for *within-condition variance* as well. In the results below, we also show that the attention mechanism itself requires no reliance on this parameter for its effectiveness—it still generates the same qualitative trends when the number of epochs is held constant. Nonetheless, the capacity to accommodate individual differences in "pondering time" exemplifies how neural-based implementations can capture not only central tendencies but also richer variance structures in human data.

7.4.4 The Attention mechanism

Having laid out the basic learning pipeline and the way we sample the number of training epochs for each subject, we now introduce the *attention mechanism* that applies to participants in the ANY CAUSE and CAUSAL SELECTION conditions of our experiment. In each condition, the observations provided to people were accompanied by an explanation that highlighted a certain subset of variables as the cause of the outcome. But only in the CAUSAL SELECTION condition these explanations corresponded to the causal selection judgments we expect people to give for these same observations. This mechanism is meant to model how explanations direct the learner's focus to specific variables, thereby modulating how strongly those variables influence the learning process.

Recall from Algorithm 7 that each simulated subject sees the same set of ten training observations in random order, for E epochs, where E is drawn from an exponential distribution. In the CAUSAL SELECTION condition, each observation is paired with exactly *one* explanation that highlights a subset of variables as causally relevant. In the ANY CAUSE condition, each observation has a *list* of possible explanations, and for each subject, we randomly pick *one* of these explanations to display—importantly, this choice remains fixed across *all* epochs for that subject. This matches the experimental setup, where participants heard the same

explanation(s) repeated whenever that observation was shown. Formally, let each observation ω be associated with an *explanation set* Ξ_ω, which contains one or more possible explanations that were presented in our experiment.

- In the CAUSAL SELECTION condition, each Ξ_ω has exactly one element: $|\Xi_\omega| = 1$, which is the unique plural or singular cause explanation associated with each observation in our experiment.
- In the ANY CAUSE condition, Ξ_ω may contain multiple candidates, randomized across subjects.
- In the OBSERVATIONS ONLY condition, $\Xi_\omega = \varnothing$ (i.e., no explanation is given).

For a given subject, if $|\Xi_\omega| > 1$ as in the ANY CAUSE condition exactly one explanation is chosen *once* (at the beginning of training) from Ξ_ω and then remains fixed for all repetitions and epochs. Denote this chosen explanation by ξ_ω. Each explanation ξ_ω is formally a function mapping each variable $x \in \{A,B,C,D\}$ to $\{+1,-1\}$, such that

$$\xi_\omega(x) \;=\; \begin{cases} +1, & \text{if variable } x \text{ is mentioned in the explanation,} \\ -1, & \text{otherwise.} \end{cases}$$

Thus, ξ_ω encodes which variables are highlighted $(+1)$ versus unmentioned (-1) in the explanation for observation ω.

To model the impact of each explanation, we then translate $\xi_\omega(x)$ into an *attention weight* via

$$\mathrm{Att}(x,\omega) \;=\; \exp\big[\alpha\,\xi_\omega(x)\big],$$

where α is an *attention parameter* shared across all observations. If x appears in the (chosen) explanation ξ_ω, then $\xi_\omega(x) = +1$, and its weight is e^{α}. If x is omitted, $\xi_\omega(x) = -1$, and its weight is $e^{-\alpha}$. As α grows positive, the difference between $e^{+\alpha}$ and $e^{-\alpha}$ increases, so the variables mentioned in explanations receive progressively more emphasis.

Incorporating Attention in the Pipeline. Recall that each variable $x \in \{A,B,C,D\}$ takes on an observed value of $+1$ (if a colored ball is drawn) or -1 (if white). Under the attention mechanism, the network's input for variable x in observation ω becomes

$$x^{\omega}_{\text{input}} \;\longleftarrow\; \big(\pm 1\big) \times \mathrm{Att}(x,\omega).$$

Hence, variables mentioned in ξ_ω are *up-weighted*, whereas unmentioned ones are *down-weighted*. If $\alpha = 0$, then $\mathrm{Att}(x,\omega) \equiv 1$, and no attention bias is applied. This recovers the OBSERVATIONS ONLY case. The ANY CAUSE and CAUSAL SELECTION conditions differ solely in which ξ_ω is chosen and how many variables it highlights—both still operate under the same α value, reflecting the intuitive assumption that explanations in either condition direct attention with the same overall strength.

7.4.5 Applying the Model to Data

With the model fully specified, we now show how it captures the core patterns of accuracy observed in our experiment. Figure 7.5 reproduces the empirical distribution of accuracy across the three conditions, reported previously. As discussed earlier, participants in AC performed worse on average than those in OBS, who

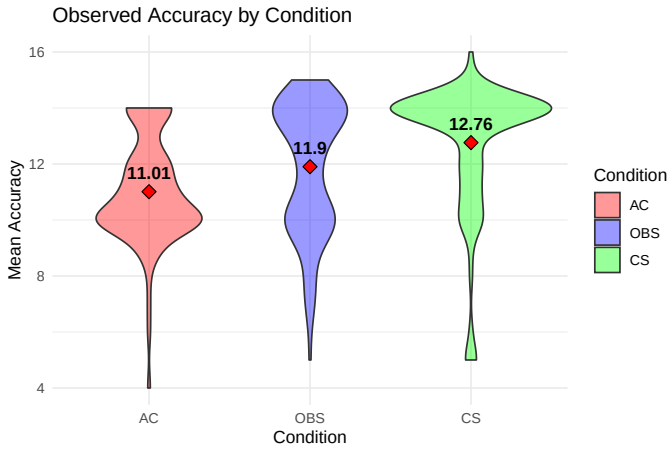


Figure 7.5: Empirical accuracy across conditions (violin plots). The *Any Cause* (red) condition yields the lowest mean accuracy, the *Causal Selection* (green) condition the highest, with *Observation-Only* (blue) in between.

in turn performed worse than those in CS. This pattern was statistically significant between every pair of conditions.

Example of model application. We first illustrate how the attention mechanism alone can reproduce these condition-wise differences without fine-tuning parameters. Specifically, we set $\alpha = 1$ and $\lambda = 1{,}000$. Here the high value $\lambda = 1{,}000$ basically makes nearly every simulated subject train for *exactly* one epoch (since $x \sim \text{Exponential}(1{,}000)$ is almost always near zero, and we take $E = \max(1, \text{round}(x))$), so that there is no difference in training times between all of the simulations generated, further zooming in on the effect of the attention mechanism. The value $\alpha = 1$ is some arbitrary moderate-to-high value taken to serve as anchor point.

Figure 7.6 shows the distribution of results for 1,000 simulated "subjects." The model recovers the correct *rank ordering* of accuracies: AC $<$ OBS $<$ CS. The absolute accuracy in each condition is higher than the human data. This phenomenon is unrelated to the attention mechanism, as evidenced by the fact that it also occurs in the observation condition (where attention plays no role). My guess is that it is probably due to the simplifications we made in building our model, a point I will return to in the discussion later on. The attention mechanism, however, is sufficient to yield the same qualitative trends seen in the experiment.

Exploring the Effect of α. To further zoom in on how the attention mechanism drives differences across conditions, we simulated the model for a range of α values (again fixing $\lambda = 1{,}000$ so that each subject trains for exactly one epoch). Figure 7.7 shows how the mean accuracy in the CAUSAL SELECTION , ANY CAUSE and OBSERVATIONS ONLY conditions changes as α varies. Three noteworthy patterns emerge in Figure 7.7:

1. **Initial benefits for both CS and AC.** When α increases from 0 to about 0.1, the accuracy of *both* the CAUSAL SELECTION (CS) and ANY CAUSE (AC) conditions actually improves relative to $\alpha = 0$. This initially mild attention bias appears helpful even in AC, since focusing on any subset of relevant

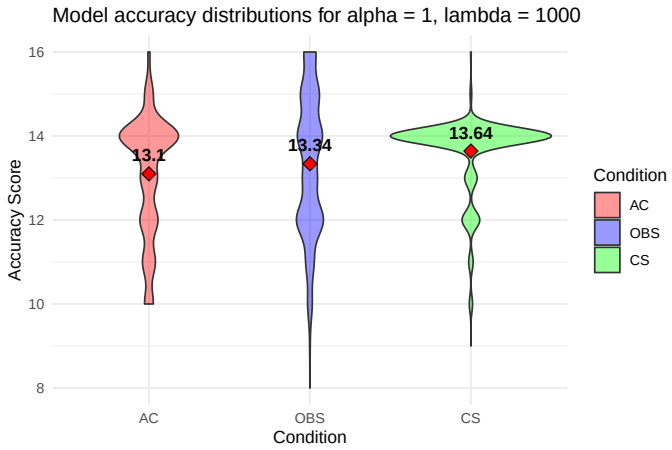


Figure 7.6: Simulated accuracy distributions for $\alpha = 1, \lambda = 1{,}000$. The ordering AC $<$ OBS $<$ CS emerges, mirroring the data, although the absolute accuracies are higher than empirical values.

variables (even if not the most important) still helps in weeding out irrelevant ones, such as urn `B` in our task.

2. **Divergence between CS and AC at higher α.** Past this small interval, further increases in α push the CS condition's performance higher while causing AC accuracy to drop. In AC, explanations can include unnecessary variables, and when α becomes too large, the learner overemphasizes these extraneous factors at the expense of the truly critical ones. Conversely, in CS, highlighting the correct variable(s) more strongly continues to boost performance until roughly $\alpha \approx 0.5$.

3. **Overshoot at very large α.** Beyond $\alpha \approx 0.5$, even the CS condition sees diminishing returns. As α grows extreme (with values of 5 and higher, not shown on the plot), the learner becomes so narrowly focused on the highlighted variable(s) that it eventually performs *worse* than the no-explanation baseline ($\alpha = 0$).

These trends make intuitive sense if we consider how explanations are supposed to direct, rather than constrain, the learner's attention. To reuse an example from before, your cousin's comment that "He won because of his strong serve" is helpful as long as it highlights the importance of a factor you may have otherwise underestimated. But suppose that it made you so single-pointedly focused on the serve that you basically start to ignore everything that happens after it. This certainly would not be ideal for learning about tennis. It may even make you misunderstand the importance of serves because you ignored the surrounding context that makes them important. Having established that our attention mechanism captures these nuances in how explanations orient the learner, we now proceed to fit α (and λ) more precisely to the empirical data.

Fitting Parameters to Data. We use a grid search over (α, λ) to minimize the sum of squared errors (SSE) between the model's pairwise condition differences (mean accuracies) and those observed in the experiment. Concretely, if we denote the empirical means by $m_{\text{OBS}}, m_{\text{AC}}, m_{\text{CS}}$ and the model-predicted means

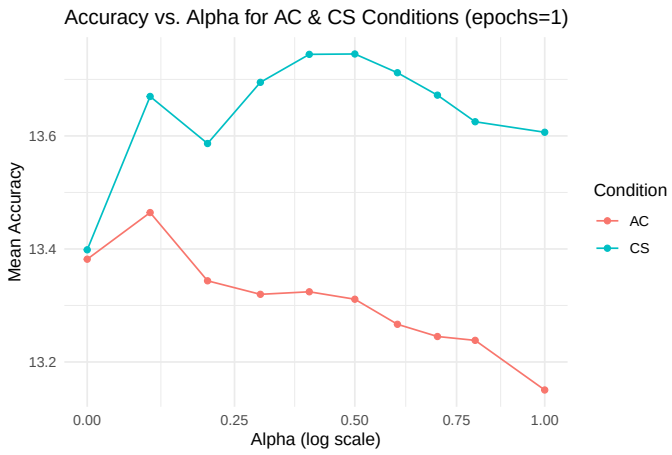


Figure 7.7: Mean accuracy as a function of the attention parameter α, with $\lambda = 1,000$ (i.e., each simulated subject completes exactly one epoch). Notice how the CS and AC conditions both initially benefit from low levels of attention, then diverge as α increases.

by $\hat{m}_{\text{OBS}}, \hat{m}_{\text{AC}}, \hat{m}_{\text{CS}}$, we define:

$$\text{SSE} = \left(\Delta_{\text{AC,OBS}} - \hat{\Delta}_{\text{AC,OBS}}\right)^2 + \left(\Delta_{\text{CS,OBS}} - \hat{\Delta}_{\text{CS,OBS}}\right)^2 + \left(\Delta_{\text{CS,AC}} - \hat{\Delta}_{\text{CS,AC}}\right)^2,$$

where $\Delta_{\text{X,Y}} = m_{\text{X}} - m_{\text{Y}}$, and similarly for $\hat{\Delta}$. The best fit yields $\alpha = 1.725$ and $\lambda = 2.35$, producing an SSE of ~ 0.6 (placeholder) and generating the predictions shown in Figure 7.8. Notably, these fitted values do not *drastically* alter the mean accuracies or differences between conditions beyond what we saw with $\alpha = 1, \lambda = 1,000$. The main difference is that with moderate λ and lower α, the model shows slightly greater variance, in particular in the CS condition, as the lower value of α did not force all of the "subjects" to converge to the same solution out of an almost exclusive reliance on the highlighted variables. The λ parameter did not make much of a contribution to the final shape of the results, compared to the previous example where the number of epochs was artificially frozen to 1 for all subjects.

7.4.6 Summary and Discussion of Results

We have compared two broad perspectives on how participants use causal explanations:

1. *A Reverse-Engineering account:* People infer causal rules by hypothesizing a generative process that could have produced a given explanation ξ, then updating rule probabilities accordingly.

2. *An Attention-Based account:* Explanations function as *instructions* that emphasize certain variables, thereby modulating how strongly those variables influence a gradient-based learning update.

Beyond its ease of *implementation* and *scalability* (as discussed in Section 7.3), the attention-based account has strong *psychological* plausibility in its own right. It aligns with the notion that explanations make it *easier* for learner to draw inferences from what they observe. In contrast, the reverse-engineering approach – or at least the version of it that is based on bayesian inference – supposes that people compute a likelihood function

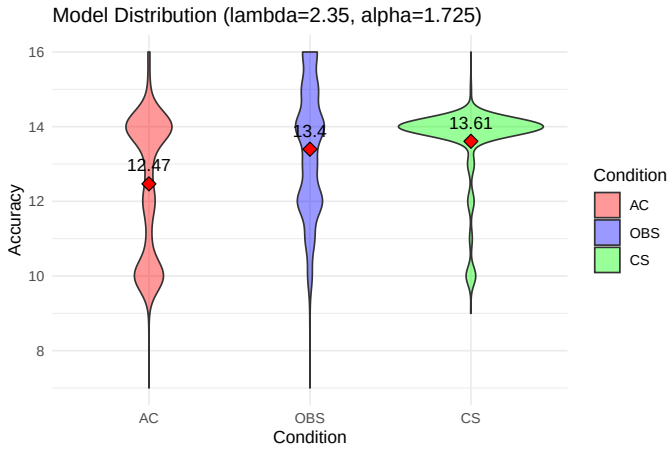


Figure 7.8: Model predictions under the best-fitting (α, λ) against empirical means (points with error bars). Note the overall closer variance match compared to the non-fitted case.

for many explanations and under many rules, which can be highly complex for larger tasks. Empirically, the experiment presented in Chapter 7 also revealed key findings that the reverse-engineering model fails to accommodate.

By contrast, a gradient-based learning model complemented with an attention mechanism readily handles these findings using minimal assumptions. It naturally explains why a single anomalous mention of `D` does not dislodge `W:= C` when the rest of the data and explanations reinforce `C`, and predicts that providing "non-causal selection" explanations misleads learners more than giving them no explanation at all in this particular task.

Despite these successes, the model we presented was intentionally minimalist and as such cannot capture every nuance of the data. In the following, we highlight several possible extensions.

Overconfidence from Simplifications. Our model often attains higher accuracies than participants do. A central reason is that we effectively set the learning rate to 1, which is unusually large by typical machine-learning standards. This makes the network *overconfident*, latching rapidly onto the first consistent pattern the training data suggests. Indeed, further tests (not reported here) reveal that using a lower learning rate brings average performance closer to that of our participants, even when we allow more epochs as compensation for the slower rate of learning. This suggests that part of the over-performance observed comes from the fact that latching onto the first consistent pattern just happens (by coïncidence) to be a good strategy in this particular task design.

Incorporating a stopping rule for learning time. We modeled response times by sampling a fixed number of epochs from an exponential distribution, after which learning abruptly stops. However, a more realistic approach might be to make the exponential distribution of epochs itself an emergent effect of the model's internal dynamics, by virtue of some *stopping criterion*. For instance, the learner could re-sample small, randomized batches from the set of observations, compute gradient updates, and repeat this process until the size of the updates fall below a certain threshold or begin to oscillate—signaling a form of saturation. This

would be analogous to a drift-diffusion process, where each new mismatch with the current rule prolongs training. Such an approach could produce exponential-like waiting times endogenously, better capturing participants' varying "pondering" durations and linking them directly to changes in the internal state of the learner.

Interaction with Explicit Hypotheses. In its current form, our model assumes that participants learn *only* through incremental parameter updates in an implicit manner. In everyday reasoning, however, people sometimes pause to articulate a symbolic rule for themselves—for instance, concluding that `W:= C` is consistent with their accumulated observations. A *hybrid* neural-symbolic framework (e.g., Garcez, Broda, and Gabbay 2002) could accommodate such behavior by periodically translating the network's learned parameters into an explicit logical program, using established *knowledge extraction procedures*. Crucially, the extracted rule might not perfectly reproduce the network's function—small mismatches can arise because the network's weight-based decisions can be more fine-grained than the symbolic rules they induce. Studying how predictions change when learners are prompted to articulate their beliefs explicitly (e.g., by writing out the rule before generating predictions) would shed light on how this *explicit* articulation reshapes subsequent inferences.

Despite the simplifications used, the current model demonstrates the core viability of the attention-based approach. By highlighting how explanations shape learners' focus, it captures not only the overall ordering of conditions but also many subtle phenomena that purely reverse-engineering explanations cannot accommodate.

7.5 General conclusion

In this dissertation, I sought out to understand how discrete "yes-or-no" notions of causality coexist in human minds with the intuition that different causes contribute *more* than others to an outcome. I studied how people assign causal responsibility to events, and how these assignments manifest themselves in the explanations they offer for why certain outcomes happen. I looked both at how our responsibility assignments depend on the shape of our causal knowledge, and how causal explanations we are given themselves inform our causal knowledge. A common thread to my argument in both cases was that causal responsibility should be understood as mapping directly onto the continuous weight parameters of an internal model of causal relations.

By way of conclusion, I would like to briefly consider how the scope of such an account might be extended to the distinct but related domain of probabilistic judgment. Specifically, I would like to consider how it might resonate with the broader research program that goes under the heading of *Bayesian confirmation theory*. The core idea is as follows. When faced with an inference problem involving a set of hypotheses $\mathscr{H}$ and a body of evidence e , people often adjudicate among hypotheses $h \in \mathscr{H}$ by relying on the extent to which each hypothesis is *confirmed* by e, rather than on the posterior probability of each hypothesis given e. The extent in which an hypothesis is confirmed by some piece of evidence is typically understood via measures of *evidential weight* that track the extent to which learning e increases the probability of h, using such measures as the likelihood ratio or Bayes factor:

$$\ell(H,E) \;=\; \ln \frac{\mathbb{P}(E \mid H)}{\mathbb{P}(E \mid \neg H)},$$

Confirmation-theoretic measures offer a great descriptive paradigm to capture human inference in a variety of

tasks.(Crupi, Fitelson, and Tentori 2008; Powell and Nair 2023; Sablé-Meyer and Mascarenhas 2021; Tentori, Crupi, and Russo 2013) In particular, it does a great job of capturing a number of patterns of fallacious inferences, where people draw conclusions not in line with the standards of Bayesian rationality. In spite of its successes as a descriptive paradigm, the approach raises a key concern, exacerbated by the fact that *all* major confirmation measures are spelled out in terms of conditional probabilities, like the ℓ measure presented above. **If** we take this notation literally, as an indicator of the sort of information that people explicitly combine as they draw inferences, this would suggest that to compute information metrics like ℓ, people already need to be able to represent the full joint probability distribution over the variables of a domain. In which case it becomes puzzling that people's judgment would diverge from the standards of bayesian rationality; In separate work (Konuk, Navarre, and Mascarenhas 2023) (not included in this dissertation for brevity), we showed that people's preference for confirmation-based reasoning is amplified when they can interpret the link between hypothesis and evidence in explicitly *causal* or *explanatory* terms. This suggest extending the thesis that causal responsibility assignments correspond to internal weight parameters beyond causality itself and into the realm of broader probabilistic inference, under the idea that people may be tracking confirmation relations precisely because those, much like explanatory relevance and unlike joint probability distributions, can be read directly off of the parameters of an internal model. This notion intersects with the fact that neurally plausible models for encoding probabilistic relations between variables tend to see those as emergent over a network of pairwise connection weights, where each weight can be interpreted as a confirmation measure relating the corresponding two variables.[5] This gains us a unified account of both causal and probabilistic inferences, revealing how key judgments of "why" and "how likely" can emerge from one consistent underlying architecture.

[5] In particular it can be shown that in a Boltzmann machines encoding a probability distribution $\mathbb{P}$, the pairwise connection weight $w_{A,B}$ between two neurons A and B can be rewritten as the sum of two log-likelihood ratio relations involving A and B. This follows from previously observed equivalence between $w_{A,B}$ and the ratio of pairwise marginal probability distribution over A,B observed by Hinton and Sejnowski (1983). Readers interested can find a derivation in Appendix Chapter B.

Input:

- (I, O): an observation specifying input variables I and outcomes O.
- ξ: an explanation highlighting one or more variables in I.
- N_0: a feedforward network initialized with parameters Γ_0, with input layer V_{in} and output layer V_{out}

Output: An updated network N_n with parameters Γ_n that fit both the observed outcome data and the explanatory constraints.

Initialization: Initialize N_0 with parameters Γ_0, **then do:**

1: **Step 1:** Set the activations of input neurons in V_{in} to match the values of variables in I (e.g. $+1$ for events that occurred, -1 for events that did not).

2: **Step 2: Outcome Targets** Set the activation targets of the output layer to match the the values of variables in O.

3: **Step 3 (New): Explanation Targets** Assign a causal importance target κ^* for the variable(s) C mentioned in ξ, setting some desired importance level.

4: **Step 4: Forward Pass:** Propagate the input activations through the network to obtain predicted outputs and compute the current network's relevance score for the highlighted variables.

5: **Step 5 (Modified Loss): Compute a composite loss measure for each $\gamma \in \Gamma$ as:**

$$\mathscr{L}(\gamma) \;=\; \mathscr{L}_{\omega}(\gamma) \;+\; \mathscr{L}_{\xi}(\gamma),$$

where:

- $\mathscr{L}_{\omega}(\gamma)$ measures predictive error (e.g. mean-squared error) relative to the distance between outcome targets and their effective outcome values,
- $\mathscr{L}_{\xi}(\gamma)$ measures explanatory error, penalizing deviation of the score $\kappa(C, \omega)$ from the target κ^*.

6: **Step 6 (Combined Backprop): Parameter Updates** Perform backpropagation on $\mathscr{L}(\gamma)$ for each $\gamma \in \Gamma$, combining gradients from both $\mathscr{L}_{\omega}$ and $\mathscr{L}_{\xi}$, then using them to perform updates proportional to the gradients. Each may be assign a different learning rate.

$$\nabla_{\gamma}\mathscr{L}(\gamma) = \nabla_{\gamma}\mathscr{L}_{\omega}(\gamma) \;+\; \nabla_{\gamma}\mathscr{L}_{\xi}(\gamma).$$

7: **Step 7: Iterate:** Update all parameters γ to reduce $\mathscr{L}(\gamma)$, and repeat Steps 1–6 until convergence or until a stopping criterion is reached.

Algorithm 5: Enriched Gradient Descent with Explanations

Input: $\omega = (I, O)$: an observation with input I and outcome O,
ξ: an explanation identifying a relevant subset C of I,
N_0: a feedforward network (as in Algorithm 4) with parameters γ_0.
Output: An updated network that has been guided to focus on C.

Step 1: Initialize N_0 with parameters Γ_0;
Step 2: Apply an attention mask $\alpha(\xi)$ to the input layer by multiplying the activations of variables in C by a factor > 1 and/or those not in C by factor < 1;
Step 3: Run a forward pass through the network and compute the predictive loss (as in Algorithm 4);
Step 4: Run backpropagation to update each $\gamma \in \Gamma$ in the direction that minimizes the loss;
Step 5: Repeat Steps 2–4 with various observations ω and explanations ξ as available.

Algorithm 6: Attention-Based Inductive Learning

Input: Set of 10 training observations (with repeated draws), each an input-output pair $(\mathbf{x}, y)$,
16 test configurations $(\mathbf{x}_{test}, y_{test})$, rate parameter λ.
Output: Accuracy distribution across N simulated subjects.

for $i \leftarrow 1$ **to** N **do**

  Step 1.1: Sample $x \sim \text{Exponential}(\lambda)$.
  Step 1.2: Set $E \leftarrow \max(1, \text{round}(x))$.
  Step 1.3: Initialize a network N_i with small random values as weights (see Figure 7.3).

  Step 2: for *epoch* $\leftarrow 1$ **to** E **do**
    foreach *training observation* $(\mathbf{x}, y)$ **do**
      Run a forward pass on N_i with input $\mathbf{x}$.
      Compute error (Mean Squared Error or Cross-Entropy).
      Backpropagate error and update parameters.
    end
  end

  Step 3: foreach *test configuration* $(\mathbf{x}_{test}, y_{test})$ **do**
    $\hat{y}_{test} \leftarrow \text{forwardPass}(\mathbf{x}_{test})$.
    if $\hat{y}_{test} > 0$ **then**
      prediction $\leftarrow +1$.
    end
    else
      prediction $\leftarrow -1$.
    end
    Record whether prediction $= y_{test}$.
  end

  Step 4: Compute $\text{accuracy}_i = \frac{\#\{\text{correct predictions}\}}{16}$.

end

Algorithm 7: Simulation Pipeline for N Virtual Subjects

Appendices

Appendix A

Logic Programs — Definitions and proof sketch of the equivalence with SCMs

A.1 Definitions and proof sketch of the equivalence with SCMs

A.2 Semantics of Logic Programs (Definite Case)

Definition 8 (Herbrand Base)**.** *Given a (propositional) logic program P, the* Herbrand base $\mathscr{B}_P$ *is the set of all propositional atoms appearing in P.*

Definition 9 (Interpretation)**.** *An* interpretation *I for a logic program P is a mapping from each atom in $\mathscr{B}_P$ to a truth value in $\{\top, \bot\}$. Equivalently, we identify I with the subset of $\mathscr{B}_P$ that it maps to $\top$.*

Definition 10 (Herbrand Model)**.** *A* Herbrand model *of P is an interpretation $M \subseteq \mathscr{B}_P$ such that every clause*

$$A_0 \leftarrow A_1, \ldots, A_n$$

in P is satisfied by M. For definite programs, this means: if $A_1, \ldots, A_n \in M$, then $A_0 \in M$.

Definition 11 (Least Herbrand Model)**.** *For a* definite *logic program P, there is a unique minimal (w.r.t. set inclusion) Herbrand model M_P that is contained in every other model of P. This M_P is called the* least Herbrand model *of P.*

Remark 1. The least Herbrand model M_P represents the smallest set of atoms that must be true for P to hold. Intuitively, it represents the set of "sure things" for a definite logic program, with no extraneous assumptions. We will see that when extended to General Logic Programs, it will represent a fundamental semantic anchor for proving Soundness and Completeness for those programs.

A.2.1 The Immediate Consequence Operator and the Least Fixpoint

Definition 12 (Immediate Consequence Operator T_P)**.** *For a definite logic program P, define $T_P : \mathscr{P}(\mathscr{B}_P) \to \mathscr{P}(\mathscr{B}_P)$ as:*

$$T_P(I) = \{A_0 \mid \exists (A_0 \leftarrow A_1, \ldots, A_n) \in P \text{ with } A_1, \ldots, A_n \in I\}.$$

$T_P(I)$ captures the new atoms that become derivable if we assume all atoms in I are true.

Remark 2. One can apply T_P iteratively, starting from the empty set, builds up the consequences of P step by step:

$$T_P^0(\emptyset) = \emptyset, \quad T_P^1(\emptyset) = T_P(\emptyset), \quad T_P^2(\emptyset) = T_P(T_P(\emptyset)), \text{ and so forth.}$$

$T_P^1(\emptyset)$ will make true all of the heads that are bound to an empty body in the Program, in other words all of the facts of the programs. Then iterative application moves from those to their immediate consequences, then to the consequences of those, and so on.

Definition 13 (Least Fixpoint and Least Herbrand Model via T_P)**.** *Since T_P is monotone, the sequence*

$$\emptyset \subseteq T_P(\emptyset) \subseteq T_P^2(\emptyset) \subseteq \cdots$$

is increasing and it can be proven (by what's known as the Knaster–Tarski fixpoint theorem) that it converges to a least fixpoint:

$$M_P = \bigcup_{k=0}^{\infty} T_P^k(\emptyset).$$

This M_P is also known to exactly align with the least Herbrand model of P, for any P.

Remark 3 (Intuition for the Least Fixpoint)**.** Intuitively, what this means is that as we iteratively apply T_P as described above, we eventually reach a stable point (a fixpoint) where no new atoms can be derived. This stable point is the minimal model that contains all truths forced by P.

Example 1. Consider a definite program:

$$p \leftarrow q$$
$$q \leftarrow$$

Here, $\mathscr{B}_P = \{p, q\}$. Starting from $\emptyset$, we apply T_P:

$$T_P(\emptyset) = \{q\} \text{ because } q \leftarrow \text{ is a fact.}$$

Now apply T_P again:

$$T_P(\{q\}) = \{q, p\} \text{ since } p \leftarrow q \text{ and } q \in \{q\}.$$

The next iteration:

$$T_P(\{q, p\}) = \{q, p\}.$$

So the least fixpoint is $M_P = \{p, q\}$, which is the least Herbrand model. It includes exactly the atoms that must be true.

A.2.2 Soundness and Completeness

This notion of Least Herbrand model serves as as reference to prove Soudness and Completeness theorems for proof procedures like SLD-resolution defined over Definite Logic Programs:

- *Sound:* Anything proven is true in the least Herbrand model.
- *Complete:* Anything true in the least Herbrand model can be proven.

The proof involves showing that a procedure like SLD-resolution will eventually converge to the same fixpoint as T_P, which is known to be M_P. Thus, the least Herbrand model provides a semantic reference for aligning model-theoretic and proof-theoretic views on Definite Logic Programs.

Now, before we can use that semantic reference for our own project of proving a semantic equivalence between Logic Programs and Structural Causal Models, we should see how it extends to the class of General Logic Programs, which include *negation as failure*:

A.3 From Definite to General Logic Programs

Definition 14 (General Logic Program (GLP))**.** *A* general logic program *is like a definite program, but clauses may contain* default negation *(negation as failure), written as* $\sim A$*. A general clause has the form:*

$$A_0 \leftarrow A_1, \ldots, A_m, \sim A_{m+1}, \ldots, \sim A_n.$$

Remark 4. General Logic Programs are *non-monotonic*: the truth of some atom may depend on the absence of a proof for another atom. This can lead to multiple stable solutions or none, unlike the definite case where the least Herbrand model is unique and well-defined. To track down those solutions, one should use the following transformation:

Definition 15 (Gelfond–Lifschitz (GL) Transformation)**.** *Given a GLP P and an interpretation I, define* P^I *by:*

(i) Removing any rule containing $\sim A$ *in its body if* $A \in I$*.*

(ii) Removing all $\sim A$ *literals from the remaining rules.*

The result P^I *is a* definite *program.*

Definition 16 (Stable Model)**.** *An interpretation I is a* stable modell *of P if I is the least Herbrand model of* P^I*.*

Remark 5 (Intuitive explanation for the plurality of stable models)**.** Unlike a definite program, which has a single least Herbrand model, a GLP can have multiple stable models. This is because different choices of I can lead to different reduced programs P^I. Each P^I is definite and thus has a unique least Herbrand model, but different I's may yield different models or none at all.

In other words, the stable models arise from "candidate" interpretations I. Different candidates I can sometimes work out, giving rise to multiple stable models.

Example 2. Consider the simpel GLP:

$$a \leftarrow \sim b$$
$$b \leftarrow \sim a$$

Check $I_1 = \{a\}$:

$$P^{I_1} = \{a \leftarrow\} \text{ (since we remove the clause for } b \text{ due to } \sim a \text{ and } a \in I_1)$$

The least Herbrand model of P^{I_1} is indeed $\{a\}$, so I_1 is stable.

Check $I_2 = \{b\}$:

$$P^{I_2} = \{b \leftarrow\}$$

The least Herbrand model of P^{I_2} is $\{b\}$, so I_2 is also stable.

Here we have two stable models: $\{a\}$ and $\{b\}$. This shows how non-monotonicity can produce multiple stable solutions. Note still that other candidates, such as $\{a,b\}$ or $\{\emptyset\}$, wouldn't be stable in this model.

Intermediate summary. We have now spelled out some basic semantic notions for Logic Programs. We explained how SLD-resolution is Sound and Complete wrt. the Least Herbrandt Model of a Logic Program, and show how Stable Models extend that notion to General Logic Programs. For us to establish an equivalence between General Logic Programs and (deterministic, boolean) SCMs will involve showing that for any such SCM, it is possible to systematically construct a GLP whose stable models identify with the solutions to the SCMs. But before we can do that, we want to pin down exactly the notion of solutions for an SCM, which we do in the next section.

A.4 Structural Causal Models (SCMs)

Definition 17 (Structural Causal Model (SCM))**.** *A* deterministic, boolean SCM $\mathscr{M}$ *consists of:*

- *A set of endogenous variables* $V = \{V_1, \ldots, V_n\}$*, each* $\in \{\top, \bot\}$.
- *A set of exogenous variables* $U = \{U_1, \ldots, U_m\}$*, each fixed to a known boolean value.*
- *A set of structural equations:*
$$V_i := f_i(PA_i, U_i),$$
where $PA_i \subseteq V \setminus \{V_i\}$ *are the parents of* V_i*, and* f_i *is a boolean function.*

A solution *to an SCM is an assignment of values to all* V_i *that simultaneously satisfies all equations given the fixed* U.

Definition 18 (Valuations over variables in an SCM)**.** *A* valuation *is a function* $\nu : V \cup U \to \{\top, \bot\}$ *that assigns truth values to each variable in* $V \cup U$*. It is equivalent to the notion of* Interpretation *defined previously for Logic Programs.*

Just as we did for Intepretations, we can also identify ν *with the subset of* $V \cup U$ *that it maps to* $\top$.

Remark 6. Because we are assuming that exogenous variables already come fixed to a known boolean value (as we are only considering deterministic SCMs), we can already tell that the only valuations compatible with an SCM $\mathscr{M}$ will be those that map variables in U to the value already assign to them in $\mathscr{M}$. So for our purposes the notion is mainly about assigning values to V

Definition 19 (SCM Solution)**.** *A* solution *to* $\mathscr{M}$ *is a valuation* ν *that is:*

- *Consistent with the fixed values assigned to* U *in* $\mathscr{M}$.
- *Such that for every equation* $V_i := f_i(PA_i, U_i)$*, the following holds:*
$$\nu(V_i) = f_i(\{\nu(V_j) \mid V_j \in PA_i\}, \{U_k = \top/\bot \text{ as given}\}).$$

In other words, the valuation ν *makes each* V_i *agree with* $f_i(PA_i, U_i)$*. We do not consider variables outside* $V \cup U$*, so the notion of solution is inherently local to the SCM's set of variables.*

Remark 7 (Acyclicity not required)**.** In deterministic, boolean SCMs, if the graph of parent relations PA_i is acyclic, a unique solution often emerges (once U is fixed). SCMs are typically required to be acyclic, but we won't need to assume that for the present purposes, so the claims we will make next also generalize to the case of cyclic SCMs.

Remark 8 (Simple variables on the LHS)**.** Typically, an SCM is defined so that each equation is of the form:

$$V_i := f_i(\text{PA}_i, U_i),$$

with V_i on the left-hand side (LHS) as a distinct, named variable. This form ensures a functional relationship from parents to the child variable. So negations or other logical operations are usually confined to the RHS inside f_i. Thus an "equation" like:

$$\neg V_i := f_i(\text{PA}_i, U_i),$$

would not match the standard functional form of an SCM equation. This detail emphasizes a similarly between the structure of SCM equations and LP clauses which will help help keep the translation straightforward .

Remark 9 (Trivial propositions)**.** An SCM $\mathscr{M}$ can in principle contain trivial propositions, such as:

$$E := A \wedge (B \vee \neg B).$$

Trivial propositions like this won't however fundamentally alter the shape of the set of solutions to $\mathscr{M}$. This is because B in that example is either an exogenous or an endogenous variable. So it is either fixed to a value $\{\top, \bot\}$ or assigned a value by some other equation in $\mathscr{M}$.

A.5 Proof sketch

Claim: *For any deterministic, boolean SCM with fixed exogenous variables, there exists a General Logic Program (constructed by a systematic translation procedure described below) such that the stable models of this GLP are in one-to-one correspondence with the solutions of the SCM.*

A.5.1 Translation algorithm

Definition 20 (Translation from SCM to GLP)**.** *Suppose we have a deterministic, boolean SCM $\mathscr{M}$ with:*

- *Endogenous variables $V = \{V_1, \ldots, V_n\}$.*
- *Exogenous variables $U = \{U_1, \ldots, U_m\}$, each fixed to $\top$ or $\bot$.*
- *Structural equations $V_i := f_i(PA_i, U_i)$ for $i = 1, \ldots, n$, where f_i is a total boolean function of its inputs.*

The translation proceeds as follows:

(1) ***Exogenous facts:*** *For each exogenous variable U_j fixed to $\top$, add the fact:*

$$U_j \leftarrow$$

If an exogenous variable U_j is fixed to $\bot$, do not add a fact for it. This leaves its truth value to be inferred by default negation as needed.

(2) ***Normalizing the Boolean Functions:*** *Convert each boolean function $f_i(PA_i, U_i)$ into a normal form such as Disjunctive Normal Form (DNF). This yields an expression:*

$$f_i(PA_i, U_i) \equiv D_1 \vee D_2 \vee \cdots \vee D_k,$$

where each D_r is a conjunction of literals (an atom or its negation) over variables in $(PA_i \cup U_i)$.

(3) ***Clause Generation:*** *For each disjunct D_r that would make f_i true, create a clause defining V_i. Specifically:*

$$V_i \leftarrow B_1, B_2, \ldots, B_\ell, \sim D_1, \sim D_2, \ldots, \sim D_t$$

where:

- *$B_1, \ldots, B_\ell$ are those variables that must be true for D_r to hold (positive literals in D_r).*
- *$D_1, \ldots, D_t$ are those variables that must be false for D_r to hold (negated literals in D_r are represented as default negation $\sim D_j$ in the program).*

The resulting GLP P represents the conditions under which each V_i can be derived, mirroring exactly the conditions imposed by $f_i(PA_i, U_i)$ in the SCM.

A.5.2 The stable models of $P \implies$ Solutions to $\mathscr{M}$

Statement: If M is a stable model of the GLP P derived from an SCM $\mathscr{M}$, then the valuation defined by M on $V \cup U$ is a solution to $\mathscr{M}$.

Steps:

- By definition, M being stable means M is the least Herbrand model of P^M, the GL-reduct of P with respect to M.
- The GL-reduct P^M removes clauses that depend negatively on atoms in M, and removes all negations from the remaining clauses. Therefore, P^M is a definite logic program whose clauses are directly derived from the boolean functions f_i but evaluated under the assumption that M is correct.
- As M is the least model of P^M, for each V_i, the conditions encoded in f_i (transformed into clauses) are satisfied. If they were not, then M could not be a model of P^M.
- Since f_i precisely capture the dependencies of V_i on (PA_i, U_i), the valuation induced by M on V_i matches the functional equation $V_i := f_i(\mathrm{PA}_i, U_i)$. Thus, M enforces the same conditions as a solution of the SCM.
- Hence the interpretation given by M on V (and U) is indeed a solution to the SCM.

A.5.3 Solutions to $\mathscr{M} \implies$ Stable models of P

Statement: If ν is a solution to the SCM $\mathscr{M}$, then there exists a stable model M of the corresponding GLP P such that M encodes ν.

Steps:

- Start from a solution ν of $\mathscr{M}$. Let $M_\nu = \{V_i \mid \nu(V_i) = \top\} \cup \{U_j \mid U_j = \top\}$ be the interpretation corresponding to ν.

- Consider the GL-reduct P^{M_v}. In P^{M_v}, any clause that required a negatively stated atom which is in M_v is removed, and all default negations are eliminated. The result is a definite program whose clauses mirror the conditions of f_i, but checked under M_v.
- Because v is a solution to the SCM, the assignment in M_v meets all the f_i conditions. Thus, M_v satisfies all clauses in P^{M_v}.
- Moreover, M_v is minimal with respect to these conditions. If we could remove an atom from M_v and still have a model, that would imply v was not a correct solution to the SCM's equations (since some condition would be unnecessarily fulfilled).
- By the properties of definite programs and the least Herbrand model construction, M_v must then be the least model of P^{M_v}. Therefore, M_v is stable.

Intuitively, any SCM solution can be "replayed" as a stable model by constructing the reduct and showing that the given solution satisfies and minimally justifies all clauses. The minimality of the stable model corresponds exactly to the functional necessity of each truth assignment in the SCM solution.

A.5.4 Examples

Example 1 (Disjunction in the SCM): Consider an SCM:

$$\begin{aligned} W &= \bot \\ U &= \top \\ V_1 &:= W \\ V_2 &:= U \vee V_1 \end{aligned}$$

Here, U is exogenous and fixed to $\top$, and W is exogenous and fixed to $\bot$. Therefore:

$$\begin{aligned} V_1 := \bot \quad &\Rightarrow V_1 = \bot \\ V_2 := \top \vee \bot \quad &\Rightarrow V_2 = \top. \end{aligned}$$

Hence the unique solution to this SCM is

$$v = \{U, V_2\},$$

with $W = \bot$, $V_1 = \bot$ implicitly understood from the absence of these variables in v.

The translation to a GLP P yields:

$$\begin{aligned} U &\leftarrow \\ V_1 &\leftarrow W \\ V_2 &\leftarrow U \\ V_2 &\leftarrow V_1 \end{aligned}$$

(- No fact for W since $W = \bot$.)

Consider $M_v = \{U, V_2\}$, corresponding to the SCM solution. Apply the Gelfond-Lifschitz reduction on P with respect to M_v, we get P^{M_v}:

$$U \leftarrow$$
$$V_2 \leftarrow$$

In P^{M_v}, M_v is clearly the Least Herbrand model, matching the unique SCM solution exactly.

Example 2 (Negative Fact): Consider another SCM:

$$Z = \bot$$
$$Y := \neg Z$$
$$X := Z \wedge Y$$

Since $Z = \bot$, we have:

$$Y := \neg\bot = \top,$$

and

$$X := \bot \wedge \top = \bot.$$

Thus the unique solution is

$$v = \{Y\},$$

with $Z = \bot$ and $X = \bot$.

The translation to GLP P is as follows:

$$Y \leftarrow\sim Z$$
$$X \leftarrow Z, Y$$

Applying GL GL-reduction with the model $M_v = \{Y\}$, we get P^{M_v}:

$$P^{M_v} = \{Y \leftarrow\}$$

Again the stable model $M_v = \{Y\}$ matches exactly the SCM solution.

Example 3 (Cyclic SCM): Consider a cyclic SCM:

$$U = \top$$
$$X := U \wedge \neg Y$$
$$Y := \neg X$$

This SCM has two solutions:

1. Assign $X = \top$. If $X = \top$, then $Y = \neg X = \bot$. Since $U = \top$, $X = \top$ is consistent, so $(U = \top, X = \top, Y = \bot)$ is a solution.
2. Assign $X = \bot$. If $X = \bot$, then $Y = \neg X = \top$. And if $Y = \top$, to check consistency: $X := U \wedge \neg Y = \top \wedge \neg\top = \bot$ is also consistent. So $(U = \top, X = \bot, Y = \top)$ is another solution.

We have two solutions: $v_1 = \{U, X\}$ and $v_2 = \{U, Y\}$. We want to check that those solutions exactly overlap with the stables models of the corresponding GLP P:

$$U \leftarrow$$

$$X \leftarrow U, \sim Y$$

$$Y \leftarrow \sim X$$

Check $v_1 = \{U, X\}$. Under $M_{v_1} = \{U, X\}$, the reduct $P^{M_{v_1}}$ would essentially confirm X and not allow Y. M_{v_1} is minimal and stable, reflecting the first SCM solution.

Check $v_2 = \{U, Y\}$. Similarly, $P^{M_{v_2}}$ would confirm Y and not allow X. M_{v_2} is also minimal and stable, corresponding to the second solution.

Thus, even in this cyclic case, we have a perfect correspondence between SCM solutions and stable models.

Appendix B

Weights of a Boltzmann machine as a confirmation statistic

Weights as ratios of pairwise marginals

Hinton & Sejnowski showed that the strength of a symmetric connection (weight) between two neurons corresponds to the logarithm of an odds ratio derived from their pairwise marginal probabilities in the distribution over states of the machine. Formally, they derive that a consistent Bayesian choice for the weight w_{ij} between two binary units i and j in a Boltzmann machine is:

$$\boxed{w_{ij} = \ln \frac{P(x_i = 1, x_j = 1)\, P(x_i = 0, x_j = 0)}{P(x_i = 1, x_j = 0)\, P(x_i = 0, x_j = 1)}}$$

Here we want to show that want to show that this is equal to the sum of log likelihood ratios:

$$\boxed{\ln \frac{P(x_i = 1 \mid x_j = 1)}{P(x_i = 1 \mid x_j = 0)} + \ln \frac{P(x_i = 0 \mid x_j = 0)}{P(x_i = 0 \mid x_j = 1)}}.$$

B.0.1 Expanding the RHS

$$\begin{aligned} \text{RHS} &= \ln\left(\frac{P(x_i = 1 \mid x_j = 1)}{P(x_i = 1 \mid x_j = 0)}\right) + \ln\left(\frac{P(x_i = 0 \mid x_j = 0)}{P(x_i = 0 \mid x_j = 1)}\right) \\ &= \ln\left[\frac{P(x_i = 1 \mid x_j = 1)}{P(x_i = 1 \mid x_j = 0)} \times \frac{P(x_i = 0 \mid x_j = 0)}{P(x_i = 0 \mid x_j = 1)}\right]. \end{aligned}$$

B.0.2 Rewriting conditional probabilities in terms of joint and marginal probabilities

Recall that $P(x_i \mid x_j) = \frac{P(x_i, x_j)}{P(x_j)}$. Hence,

$$\frac{P(x_i=1 \mid x_j=1)}{P(x_i=1 \mid x_j=0)} = \frac{\frac{P(x_i=1,x_j=1)}{P(x_j=1)}}{\frac{P(x_i=1,x_j=0)}{P(x_j=0)}} = \frac{P(x_i=1,x_j=1)\,P(x_j=0)}{P(x_i=1,x_j=0)\,P(x_j=1)},$$

and

$$\frac{P(x_i=0 \mid x_j=0)}{P(x_i=0 \mid x_j=1)} = \frac{\frac{P(x_i=0,x_j=0)}{P(x_j=0)}}{\frac{P(x_i=0,x_j=1)}{P(x_j=1)}} = \frac{P(x_i=0,x_j=0)\,P(x_j=1)}{P(x_i=0,x_j=1)\,P(x_j=0)}.$$

B.0.3 Multiplying fractions

$$\begin{aligned}
&\frac{P(x_i=1,x_j=1)\,P(x_j=0)}{P(x_i=1,x_j=0)\,P(x_j=1)} \times \frac{P(x_i=0,x_j=0)\,P(x_j=1)}{P(x_i=0,x_j=1)\,P(x_j=0)} \\
&= \frac{\big[P(x_i=1,x_j=1)\,P(x_j=0)\big]\,\big[P(x_i=0,x_j=0)\,P(x_j=1)\big]}{\big[P(x_i=1,x_j=0)\,P(x_j=1)\big]\,\big[P(x_i=0,x_j=1)\,P(x_j=0)\big]} \\
&= \frac{P(x_i=1,x_j=1)\,P(x_i=0,x_j=0)}{P(x_i=1,x_j=0)\,P(x_i=0,x_j=1)}.
\end{aligned}$$

Here $P(x_j=0)$ and $P(x_j=1)$ cancel each other out in the numerator and the denominator. Thus, we get

$$\text{RHS} = \ln\Big(\frac{P(x_i=1,x_j=1)\,P(x_i=0,x_j=0)}{P(x_i=1,x_j=0)\,P(x_i=0,x_j=1)}\Big),$$

proving the claim.

Bibliography

Arendt, Hannah (1987). "Collective responsibility." In: *Amor Mundi: Explorations in the Faith and Thought of Hannah Arendt*. Ed. by James W. S.J. Bernauer. Springer, Dordrecht, pp. 43–50. DOI: 10.1007/978-94-009-3565-5_3.

Bach, Sebastian, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek (2015). "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation." In: *PLoS ONE* 10.7, e0130140. DOI: 10.1371/journal.pone.0130140.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). "Neural machine translation by jointly learning to align and translate." In: *arXiv preprint arXiv:1409.0473*.

Bareinboim, Elias, Juan D. Correa, Duligur Ibeling, and Thomas F. Icard (2022). "On Pearl's Hierarchy and the Foundations of Causal Inference." In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. 1st ed. New York, NY, USA: Association for Computing Machinery, 507–556. ISBN: 9781450395861. URL: https://doi.org/10.1145/3501714.3501743.

Baumgartner, Michael (2008). "Regularity Theories of Causation are Inconsistent with the Causal Modelling Framework." In: *Synthese* 161.2, pp. 297–312. DOI: 10.1007/s11229-007-9151-y.

— (2013). "A Regularity Theoretic Approach to Actual Causation." In: *Causality in the Sciences*. Ed. by Phyllis Illari, Federica Russo, and Jon Williamson. Oxford: Oxford University Press, pp. 93–118.

Birch, Susan A. J. and Paul Bloom (2007). "The curse of knowledge in reasoning about false beliefs." In: *Psychological Science* 18.5, pp. 382–386. DOI: 10.1111/j.1467-9280.2007.01909.x.

Bramley, Neil R., Peter Dayan, Thomas L. Griffiths, and David A. Lagnado (2016). "Formalizing Neurath's Ship: Approximate Algorithms for Online Causal Learning." In: *CoRR* abs/1609.04212. arXiv: 1609.04212. URL: http://arxiv.org/abs/1609.04212.

Bramley, Neil R., David A. Lagnado, and Maarten Speekenbrink (2015). "Conservative forgetful scholars: How people learn causal structure through sequences of interventions." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 41.3, 708–731. ISSN: 0278-7393. DOI: 10.1037/xlm0000061. URL: http://dx.doi.org/10.1037/xlm0000061.

Bruner, Jerome S., Jacqueline J. Goodnow, and George A. Austin (1956). *A Study of Thinking*. New York, NY: John Wiley & Sons.

Byrne, Ruth M. J. (Mar. 1989). "Suppressing Valid Inferences With Conditionals." In: *Cognition* 31, pp. 61–83. DOI: 10.1016/0010-0277(89)90018-8.

— (2005). *The Rational Imagination: How People Create Alternatives to Reality*. Cambridge, MA: MIT Press.

— (2016). "Counterfactual Thought." In: *Annual Review of Psychology* 67, pp. 135–157. DOI: 10.1146/annurev-psych-122414-033249.

Carruthers, Peter (2002). "The cognitive functions of language." In: *Behavioral and Brain Sciences* 25.6, pp. 657–674. DOI: 10.1017/S0140525X02000122.

Champollion, Lucas and Manfred Krifka (2016). "Mereology." In: *The Cambridge Handbook of Formal Semantics*. Ed. by Paul Dekker and Maria Aloni. Cambridge Handbooks in Language and Linguistics. Cambridge University Press. Chap. 13, pp. 369–388. DOI: 0.1017/CBO9781139236157.014.

Chater, Nicholas and Michael Oaksford (2013). "Programs as causal models: Speculations on mental programs and mental representation." In: *Cognitive Science* 37.6, pp. 1171–1191. DOI: 10.1111/cogs.12062.

Cheng, Patricia W. (1997). "From covariation to causation: A causal power theory." In: *Psychological Review* 104.2, pp. 367–405.

Chomsky, Noam (1965). *Aspects in the Theoy of Syntax*. Cambridge, Mass: MIT Press.

Chung, WooJin, Nadine Bade, Sam Blanc-Cuenca, and Salvador Mascarenhas (2022). "Question-answer dynamics in deductive fallacies without language." In: *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*. Ed. by Jennifer Culbertson, Andrew Perfors, Hugh Rabagliati, and Veronica Ramenzoni. URL: https://escholarship.org/uc/item/9711612q.

Copley, Bridget (2020). "Events are the source of causal readings in the simplest English conditionals." In: *Conditionals - Logic, Linguistics, and Psychology.* Ed. by Stefan Kaufmann and David Over. URL: https://hal.science/hal-02431650.

Crupi, Vincenzo, Branden Fitelson, and Katya Tentori (2008). "Probability, confirmation, and the conjunction fallacy." In: *Thinking & Reasoning* 14.2, pp. 182–199. DOI: 10.1080/13546780701643406.

Cybenko, George (1989). "Approximation by superpositions of a sigmoidal function." In: *Mathematics of Control, Signals, and Systems* 2.4, pp. 303–314. DOI: 10.1007/BF02551274.

Danks, David, Thomas L. Griffiths, and Joshua B. Tenenbaum (2003). "Dynamical Causal Learning." In: *Advances in Neural Information Processing Systems 15*, pp. 67–74.

Davis, Zachary and Bob Rehder (May 2020). "A Process Model of Causal Reasoning." In: *Cognitive Science* 44. DOI: 10.1111/cogs.12839.

De Leeuw, Joshua R (2015). "jsPsych: A JavaScript library for creating behavioral experiments in a Web browser." In: *Behavior research methods* 47.1, pp. 1–12.

Ding, Yifeng, Wesley H. Holliday, and Thomas F. Icard (2021). "Logics of imprecise comparative probability." In: *International Journal of Approximate Reasoning* 132, pp. 154–180. DOI: 10.1016/j.ijar.2021.02.004.

Dretske, Fred I. (1972). "Contrastive Statements." In: *Philosophical Review* 81, pp. 411–437.

Epley, Nicholas, Boaz Keysar, Leaf Van Boven, and Thomas Gilovich (2004). "Perspective taking as egocentric anchoring and adjustment." In: *Journal of Personality and Social Psychology* 87.3, pp. 327–339. DOI: 10.1037/0022-3514.87.3.327.

Feldman, Jacob (2000). "Minimization of Boolean complexity in human concept learning." In: *Nature* 407.6804, pp. 630–633. DOI: 10.1038/35036586.

Fernbach, Philip M., Adam Darlow, and Steven A. Sloman (2011). "Asymmetries in predictive and diagnostic reasoning." In: *Journal of Experimental Psychology: General* 140.2, pp. 168–185. DOI: 10.1037/a0022100.

Fernbach, Philip M. and Bob Rehder (2013). "Cognitive shortcuts in causal inference." In: *Argument & Computation* 4.1, pp. 64–88. DOI: 10.1080/19462166.2012.682655.

Fodor, Jerry (1975). *The Language of Thought*. Harvard University Press.

— (2001). "Language, thought and compositionality." In: *Mind & Language* 16.1, pp. 1–15. DOI: 10.1111/1468-0017.00153.

— (2008). *LOT 2: The Language of Thought Revisited*. Oxford University Press. ISBN: 978-0-19-958801-5.

Fodor, Jerry A. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.

Frank, Michael C. and Noah D. Goodman (2012). "Predicting pragmatic reasoning in language games." In: *Science* 336.6084, p. 998. DOI: 10.1126/science.1218633.

Garcez, Artur S. d'Avila, Krysia Broda, and Dov M. Gabbay (2002). *Neural-Symbolic Learning Systems: Foundations and Applications*. London: Springer. ISBN: 978-1852335128.

Garcez, Artur S. d'Avila, Luiz C. Lamb, and Dov M. Gabbay (2009). *Neural-Symbolic Cognitive Reasoning*. Berlin: Springer. ISBN: 978-3642015038.

Gelfond, Michael and Vladimir Lifschitz (Aug. 1991). "Lifschitz, V.: Classical Negation in Logic Programs and Disjunctive Databases. New Generation Computing 9, 365-385." In: *New Generation Computing* 9, pp. 365–385. DOI: 10.1007/BF03037169.

Gerstenberg, Tobias, Noah D. Goodman, David A. Lagnado, and Joshua B. Tenenbaum (2021). "A counterfactual simulation model of causal judgments for physical events." In: *Psychological Review* 128.5, pp. 936–975. DOI: 10.1037/rev0000281.

Gerstenberg, Tobias and Thomas F. Icard (2020). "Expectations affect physical causation judgments." In: *Journal of Experimental Psychology: General* 149.3, p. 599.

Gerstenberg, Tobias, David Lagnado, and Ro'i Zultan (June 2023). "Making a positive difference: Criticality in groups." In: *Cognition* 238, p. 105499. DOI: 10.1016/j.cognition.2023.105499.

Gerstenberg, Tobias, Matthew F Peterson, Noah D Goodman, David A Lagnado, and Joshua B Tenenbaum (2017). "Eye-tracking causality." In: *Psychological science* 28.12, pp. 1731–1744.

Gerstenberg, Tobias and Joshua B. Tenenbaum (2017). *Intuitive theories*. Oxford University Press.

Gill, Maureen, Jonathan F. Kominsky, Thomas F. Icard, and Joshua Knobe (2022). "An interaction effect of norm violations on causal judgment." In: *Cognition* 228, p. 105183.

Goodman, Noah D. and Michael C. Frank (2016). "Pragmatic language interpretation as probabilistic inference." In: *Trends in Cognitive Sciences* 20.11, pp. 818–829. DOI: 10.1016/j.tics.2016.08.005.

Green, David W. (1998). "Explanation and connectionism: Arguments, demonstrations and analogies." In: *Mind & Language* 13, pp. 242–260.

Griffiths, Thomas and Joshua Tenenbaum (Jan. 2005). "Structure and strength in causal induction." In: *Cognitive psychology* 51, pp. 334–84. DOI: 10.1016/j.cogpsych.2005.05.004.

Groenendijk, Jeroen (2008). *Inquisitive Semantics: Two possibilities for disjunction*. Tech. rep. PP-2008-26. ILLC Prepublication. Institute for Logic, Language and Computation, University of Amsterdam.

Hall, Ned (2004). "Two concepts of causation." In.

Halpern, Joseph Y. (2015). "A modification of the Halpern-Pearl definition of causality." In: *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pp. 3022–3033.

— (2016a). *Actual Causality*. The MIT Press. DOI: 10.7551/mitpress/10809.001.0001. URL: https://doi.org/10.7551/mitpress/10809.001.0001.

— (2016b). *Actual Causality*. MIT Press.

Halpern, Joseph Y. and Judea Pearl (2005). "Causes and Explanations: A Structural-Model Approach, Part I: Causes." In: *The British Journal for the Philosophy of Science* 56.4, pp. 843–887. URL: http://www.jstor.org/stable/3541870..

— (2009). "Causes and Explanations: A Structural-Model Approach. Part II: Explanations." In: *The British Journal for the Philosophy of Science* 57.1, pp. 1–35.

Hart, H. L. A. and Tony Honoré (1985). "Causation in the Law." In.

Heathcote, Andrew, Stephen J. Popiel, and D. J. K. Mewhort (1991). "Analysis of response time distributions: An example using the Stroop task." In: *Psychological Bulletin* 109.2, pp. 340–347. DOI: 10.1037/0033-2909.109.2.340.

Henne, Paul (2023). "Experimental Metaphysics: Causation." In: *The compact compendium of experimental philosophy. De Gruyter*.

Henne, Paul, Alexandra Kulesza, Karla Perez, and Augustana Houcek (2021). "Counterfactual thinking and recency effects in causal judgment." In: *Cognition* 212. DOI: 10.1016/j.cognition.2021.104708.

Henne, Paul, Laura Niemi, Angel Pinillos, Felipe De Brigard, and Joshua Knobe (2019). "A counterfactual explanation for the action effect in causal judgment." In: *Cognition* 190, pp. 157–164. DOI: 10.1016/j.cognition.2019.05.006.

Hinton, Geoffrey E. and Terrence J. Sejnowski (1983). "Optimal Perceptual Inference." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Washington, DC, USA, pp. 448–453.

Hitchcock, Christopher (2001a). "A Tale of Two Effects." In: *The Philosophical Review* 110.3, pp. 361–396. DOI: 10.1215/00318108-110-3-361.

— (June 2001b). "The Intransitivity of Causation Revealed in Equations and Graphs." In: *The Journal of Philosophy* 98, pp. 273–299. DOI: 10.2307/2678432.

— (2007). "Prevention, Preemption, and the Principle of Sufficient Reason." In: *Philosophical Review* 116.4, pp. 495–532. DOI: 10.1215/00318108-2007-009.

— (2012). "Portable Causal Dependence: A Tale of Consilience." In: *Philosophy of Science* 79.5, pp. 942–951. ISSN: 00318248, 1539767X. URL: http://www.jstor.org/stable/10.1086/667899 (visited on 04/15/2024).

Hockett, Charles F. (1960). "The origin of speech." In: *Scientific American* 203.3, pp. 88–97. DOI: 10.1038/scientificamerican0960-88.

Holliday, Wesley H. and Thomas F. Icard (2013). "Measure semantics and qualitative semantics for epistemic modals." In: *Proceedings of SALT 23*, pp. 514–534.

Holyoak, Keith J. and Patricia W. Cheng (2011). "Causal learning and inference as a rational process: The new synthesis." In: *Annual Review of Psychology* 62, pp. 135–163.

Horton, William S. and Boaz Keysar (1996). "When do speakers take into account common ground?" In: *Cognition* 59.1, pp. 91–117. DOI: 10.1016/0010-0277(96)81418-1.

Icard, Thomas F. (Aug. 2015). "Subjective Probability as Sampling Propensity." In: *Review of Philosophy and Psychology* 7. DOI: 10.1007/s13164-015-0283-y.

— (2017). "From Programs to Causal Models." In: *Proceedings of the 21st Amsterdam Colloquium*. Ed. by Alexandre Cremers, Thom van Gessel, and Floris Roelofsen, pp. 35–44.

Icard, Thomas F., Jonathan F. Kominsky, and Joshua Knobe (2017). "Normality and actual causal strength." In: *Cognition* 161, pp. 80–93.

Johnson-Laird, Philip N. (1983a). *Mental models : towards a cognitive science of language, inference, and consciousness*. Excerpts available on Google Books. Cambridge, MA: Harvard University Press, 528 p. URL: https://hal.science/hal-00702919.

— (1983b). "Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness." In.

Kahneman, Daniel and Dale T Miller (1986). "Norm theory: Comparing reality to its alternatives." In: *Psychological review* 93.2, p. 136.

Kinney, David B and Tania Lombrozo (2024). "Building Compressed Causal Models of the World." In: *Cognitive Psychology*. URL: https://osf.io/preprints/psyarxiv/2f7x6.

Kirfel, Lara, Thomas F. Icard, and Tobias Gerstenberg (July 2022). "Inference from explanation." In: *Journal of Experimental Psychology: General* 151.7, pp. 1481–1501. DOI: 10.1037/xge0001151. URL: https://doi.org/10.1037/xge0001151.

Kirfel, Lara and David Lagnado (2021). "Changing Minds — Epistemic Interventions in Causal Reasoning." on PsyArXiv. DOI: 10.31234/osf.io/db6ms.

Knobe, Joshua and Ben Fraser (2008). "Causal Judgment and Moral Judgment: Two Experiments." In: *Moral Psychology*. Ed. by Walter Sinnott-Armstrong. MIT Press.

Kominsky, Jonathan F., Jonathan Phillips, Tobias Gerstenberg, David Lagnado, and Joshua Knobe (Feb. 2015). "Causal Superseding." In: *Cognition* 137, pp. 196 –209. DOI: 10.1016/j.cognition.2015.01.013.

Konuk, Can, Michael Goodale, Tadeg Quillien, and Salvador Mascarenhas (2023a). "Plural causes in causal judgment." In: *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*. Ed. by Micah Goldwater, Florencia K. Anggoro, Brett K. Hayes, and Desmond C. Ong, pp. 3180–3186. URL: https://escholarship.org/uc/item/0014w3r1.

— (2023b). *Plural causes in causal judgment*. DOI: 10.31234/osf.io/nuptb. URL: psyarxiv.com/nuptb.

Konuk, Can, Nicolas Navarre, and Salvador Mascarenhas (2023). "Effects of causal structure and evidential impact on probabilistic reasoning." In: *Proceedings of the 45th Annual Conference of the Cognitive Science Society*. Cognitive Science Society. URL: https://api.semanticscholar.org/CorpusID:271884246.

Konuk, Can, Tadeg Quillien, and Salvador Mascarenhas (2024). "Plural causes." Manuscript under review, available on lingbuzz. URL: https://lingbuzz.net/lingbuzz/008485.

Koralus, Philipp and Salvador Mascarenhas (2013). "The erotetic theory of reasoning: bridges between formal semantics and the psychology of deductive inference." In: *Philosophical Perspectives* 27, pp. 312–365. DOI: 10.1111/phpe.12029.

— (2018). "Illusory inferences in a question-based theory of reasoning." In: *Pragmatics, Truth, and Underspecification: Towards an Atlas of Meaning*. Ed. by Ken Turner and Laurence Horn. Vol. 34. Current Research in the Semantics/Pragmatics Interface. Leiden: Brill. Chap. 10, pp. 300–322. DOI: 10.1163/9789004365445_011.

Krasich, Kristina, Kevin O'Neill, and Felipe De Brigard (2024). "Looking at Mental Images: Eye-Tracking Mental Simulation During Retrospective Causal Judgment." In: *Cognitive Science* 48.3, e13426.

Krifka, Manfred (1996). "Pragmatic strengthening in plural predications and donkey sentences." In: *Proceedings of SALT 6*. Ed. by Teresa Galloway and Justin Spence, pp. 136–153. DOI: 10.3765/salt.v6i0.2769.

Križ, Manuel and Benjamin Spector (2021). "Interpreting plural predication: Homogeneity and non-maximality." In: *Linguistics and Philosophy* 44, pp. 1131–1178. DOI: 10.1007/s10988-020-09311-w.

Lagnado, David A., Tobias Gerstenberg, and Ro'i Zultan (July 2013). "Causal Responsibility and Counterfactuals." In: *Cognitive Science* 37.6, 1036–1073. ISSN: 1551-6709. DOI: 10.1111/cogs.12054. URL: http://dx.doi.org/10.1111/cogs.12054.

Lampinen, Andrew K., Nicholas A. Roy, Ishita Dasgupta, Stephanie C. Y. Chan, Allison C. Tam, James L. McClelland, Chen Yan, Adam Santoro, Neil C. Rabinowitz, Jane X. Wang, and Felix Hill (2021). "Tell me why! - Explanations support learning of relational and causal structure." In: *CoRR* abs/2112.03753. arXiv: 2112.03753. URL: https://arxiv.org/abs/2112.03753.

Lappin, Shalom (1989). "Donkey pronouns unbound." In: *Theoretical Linguistics* 15.3, pp. 263–289. DOI: 10.1515/thli.1988.15.3.263.

Leeuw, Joshua R. de, Rebecca A. Gilbert, and Björn Luchterhandt (2023). "jsPsych: Enabling an Open-Source Collaborative Ecosystem of Behavioral Experiments." In: *Journal of Open Source Software* 8.85, p. 5351. DOI: 10.21105/joss.05351. URL: https://doi.org/10.21105/joss.05351.

Lewis, David (1973a). "Causation." In: *The Journal of Philosophy* 70.17, pp. 556–567. DOI: 10.2307/2025310.

— (1973b). *Counterfactuals*. Oxford: Basil Blackwell.

Link, Godehard (1983). "The logical analysis of plural and mass terms: a lattice-theoretical approach." In: *Meaning, Use, and Interpretation of Language*. Ed. by R. Bäuerle, C. Schwarze, and Arnim von Stechow. Berlin: Walter de Gruyter, pp. 302–323.

Lloyd, J. W. (1984). *Foundations of Logic Programming*. Berlin: Springer-Verlag.

Löbner, Sebastian (2000). "Polarity in natural language: Predication, quantification and negation in particular and characterizing sentences." In: *Linguistics and Philosophy* 23, pp. 213–308. DOI: 10.1023/A:1005571202592.

Lombrozo, Tania (2010). "Causal-explanatory pluralism: How intensions, functions, and mechanisms influence causal ascriptions." In: *Cognitive Psychology* 61.4, pp. 302–332. DOI: 10.1016/j.cogpsych.2010.05.002.

Lombrozo, Tania and Nicholas Z. Gwynne (Sept. 2014). "Explanation and inference: mechanistic and functional explanations guide property generalization." In: *Frontiers in Human Neuroscience* 8. ISSN: 1662-5161. DOI: 10.3389/fnhum.2014.00700. URL: http://dx.doi.org/10.3389/fnhum.2014.00700.

Lu, Hongjing, Alan L. Yuille, Mimi Liljeholm, Patricia W. Cheng, and Keith J. Holyoak (2008). "Bayesian Generic Priors for Causal Learning." In: *Psychological Review* 115.4, pp. 955–984. DOI: 10.1037/a0013256.

Lucas, Christopher, Thomas Griffiths, Joseph Williams, and Michael Kalish (Mar. 2015). "A rational model of function learning." In: *Psychonomic bulletin & review* 22. DOI: 10.3758/s13423-015-0808-5.

Lucas, Christopher and Charles Kemp (2015). "An improved probabilistic account of counterfactual reasoning." In: *Psychological review* 122.4, p. 700.

Malamud, Sophia A. (2012). "The Meaning of Plural Definites: A Decision-Theoretic Approach." In: *Semantics & Pragmatics* 5.3, pp. 1–58. DOI: 10.3765/sp.5.3. URL: https://doi.org/10.3765/sp.5.3.

Mandelbaum, Eric, Yarrow Dunham, Roman Feiman, Chaz Firestone, E.J. Green, Daniel Harris, Melissa M. Kibbe, Benedek Kurdi, Myrto Mylopoulos, Joshua Shepherd, Alexis Wellwood, Nicolas Porot, and Jake Quilty-Dunn (2022). "Problems and mysteries of the many languages of thought." In: *Cognitive Science* 46.12. DOI: 10.1111/cogs.13225.

Marr, David (1982a). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman.

— (1982b). *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman and Company.

Mascarenhas, Salvador (2009). "Inquisitive Semantics and Logic." M.Sc. Thesis. University of Amsterdam.

Massaro, Dominic W. (1988). "Some criticisms of connectionist models of human performance." In: *Journal of Memory and Language* 27, pp. 213–234.

McCarthy, John and Patrick J. Hayes (1969). "Some philosophical problems from the standpoint of artificial intelligence." In: *Machine Intelligence 4*. Ed. by Bernard Meltzer and Donald Michie. Edinburgh: Edinburgh University Press, pp. 463–502.

Miller, Seumas (2001). "Collective responsibility." In: *Public Affairs Quaterly* 15.1, pp. 65–82. URL: https://www.jstor.org/stable/40441276.

Minsky, Marvin and Seymour Papert (1969). *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA, USA: MIT Press.

Montavon, Grégoire, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller (2017). "Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition." In: *Pattern Recognition* 65, pp. 211–222. DOI: 10.1016/j.patcog.2016.11.008.

Montavon, Grégoire, Wojciech Samek, and Klaus-Robert Müller (2018). "Methods for Interpreting and Understanding Deep Neural Networks." In: *Digital Signal Processing* 73, pp. 1–15.

— (2019). "Layer-Wise Relevance Propagation: An Overview." In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Ed. by Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. Vol. 11700. Lecture Notes in Computer Science. Springer, Cham, pp. 193–209. DOI: 10.1007/978-3-030-28954-6_10.

Morris, Adam, Jonathan Phillips, Tobias Gerstenberg, and Fiery Cushman (2019). "Quantitative causal selection patterns in token causation." In: *PLoS ONE* 14.8. DOI: 10.1371/journal.pone.0219704.

Morris, Adam, Jonathan Scott-Philips, Thomas F. Icard, Joshua Knobe, Tobias Gerstenberg, and Fiery. Cushman (2018). "Judgments of actual causation approximate the effectiveness of interventions." Psy ArXiv. DOI: 10.31234/osf.io/nq53z..

Nam, Andrew, Christopher Hughes, Thomas F. Icard, and Tobias Gerstenberg (May 2023). "Show and tell: Learning causal structures from observations and explanations." In: *Proceedings of the 45th Annual Conference of the Cognitive Science Society*. Center for Open Science. DOI: 10.31234/osf.io/wjs9q. URL: http://dx.doi.org/10.31234/osf.io/wjs9q.

Navarre, Nicolas, Can Konuk, Neil R. Bramley, and Salvador Mascarenhas (2024). "Functional rule inference from causal selection explanations." English. In: *Proceedings of the 46th Annual Conference of the Cognitive Science Society*. Ed. by Larissa K Samuelson, Stefan Frank, Mariya Toneva, Allyson Mackey, and Eliot Hazeltine. Vol. 46. Proceedings of the Annual Conference of the Cognitive Science Society. The 46th Annual Meeting of the Cognitive Science Society, COGSCI 2024 ; Conference date: 24-07-2024 Through 27-07-2024. United States: eScholarship University of California, pp. 1197–1203. URL: https://cognitivesciencesociety.org/cogsci-2024.

Novick, Laura R. and Patricia W. Cheng (2004). "Assessing interactive causal influence." In: *Psychological Review* 111.2, pp. 455–485. DOI: 10.1037/0033-295X.111.2.455.

O'Neill, Kevin, Paul Henne, Paul Bello, John Pearson, and Felipe De Brigard (2022). "Confidence and gradation in causal judgment." In: *Cognition* 223. DOI: 10.1016/j.cognition.2022.105036.

O'Neill, Kevin, Paul Henne, Rahel Pearson, and Felipe De Brigard (2021). "Causal judgments about atypical actions are influenced by agents' epistemic states." In: *Cognition* 206, p. 104485.

Pearl, Judea (1999). "Probabilities of Causation: Three Counterfactual Interpretations and Their Identification." In: *Synthese* 121.1-2, pp. 93–149.

— (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press. ISBN: 978-0521895606.

Pearl, Judea and Dana Mackenzie (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc. DOI: 10.5555/3238230.

Pearl, Judea and Thomas S. Verma (1991). "A Theory of Inferred Causation." In: *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference (KR-91)*. Ed. by James A. Allen, Richard Fikes, and Erik Sandewall. San Mateo, CA: Morgan Kaufmann, pp. 441–452.

Perales, José C., Andrés Catena, and Antonio Maldonado (2007). "The role of mechanism beliefs in causal reasoning." In: *Experimental Psychology* 54, pp. 269–277.

Perales, José C. and David R. Shanks (2017). "Models of covariation-based causal judgment: A review and synthesis." In: *Psychonomic Bulletin & Review* 24, pp. 1–23.

Phillips, Jonathan and Angelika Kratzer (2024). "Decomposing modal thought." In: *Psychological Review* 131.4, pp. 966–992. DOI: 10.1037/rev0000481.

Picat, Léo and Salvador Mascarenhas (2024). "On the interplay between interpretation and reasoning in compelling fallacies." In: *Cognitive Science* 48.12. DOI: 10.1111/cogs.70021.

Popper, Karl R. (1959). "The propensity interpretation of probability." In: *The British Journal for the Philosophy of Science* 10.37, pp. 25–42. URL: https://www.jstor.org/stable/685773.

Powell, Derek and Shyam Nair (2023). "Bayesian Confirmation and Commonsense Notions of Evidential Strength." In: *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*.

Quillien, Tadeg (2020). "When do we think that *X* caused *Y*?" In: *Cognition* 205. DOI: 10.1016/j.cognition.2020.104410.

Quillien, Tadeg and Michael Barlev (2022). "Causal Judgment in the Wild: Evidence from the 2020 U.S. Presidential Election." In: *Cognitive Science* 56.2. DOI: 10.1111/cogs.13101.

Quillien, Tadeg and Christopher Lucas (2023). "Counterfactuals and the logic of causal selection." In: *Psychological Review*. DOI: 10.31234/osf.io/ts76y.

Quillien, Tadeg, Aba Szollosi, Neil R Bramley, and Christopher Lucas (2023). "Causal inference shapes counterfactual plausibility." In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 45. 45.

Quilty-Dunn, Jake, Nicolas Porot, and Eric Mandelbaum (2023). "The best game in town: The re-emergence of the Language of Thought hypothesis across the cognitive sciences." In: *Behavioral and Brain Sciences* 46.e292. DOI: 10.1017/S0140525X22002849.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Ratcliff, Roger (1978). "A theory of memory retrieval." In: *Psychological Review* 85.2, pp. 59–108. DOI: 10.1037/0033-295X.85.2.59.

Rescorla, Robert A. and Allan R. Wagner (1972). "A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement." In: *Current Theory and Research in Motivation* 2, pp. 64–99.

Robinson, J. Alan (1965). "A Machine-Oriented Logic Based on the Resolution Principle." In: *Journal of the ACM* 12.1, pp. 23–41.

Rooth, Mats (1992). "A THEORY OF FOCUS INTERPRETATION." In: *Natural Language Semantics* 1.1, pp. 75–116. ISSN: 0925854X, 1572865X. URL: http://www.jstor.org/stable/23748778 (visited on 03/17/2025).

Rose, David, Eric Sievers, and Shaun Nichols (2021). "Cause and burn." In: *Cognition*.

Ross, Andrew Slavin, Michael C. Hughes, and Finale Doshi-Velez (2017). "Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations." In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. International Joint Conferences on Artificial Intelligence, pp. 2662–2670. DOI: 10.24963/ijcai.2017/371.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (1986). "Learning internal representations by error propagation." In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*. Ed. by David E. Rumelhart, James L. McClelland, and the PDP Research Group. Cambridge, MA: MIT Press, pp. 318–362.

Sablé-Meyer, Mathias, Kevin Ellis, Joshua B. Tenenbaum, and Stanislas Dehaene (2022). "A language of thought for the mental representation of geometric shapes." In: *Cognitive Psychology* 139.101527. DOI: 10.1016/j.cogpsych.2022.101527.

Sablé-Meyer, Mathias and Salvador Mascarenhas (2021). "Indirect illusory inferences from disjunction: a new bridge between deductive inference and representativeness." In: *Review of Philosophy and Psychology* 12.2. DOI: 10.1007/s13164-021-00543-8. URL: https://psyarxiv.com/pwkzm/.

Sloman, Steven A. and David A. Lagnado (2015). "Causality in thought." In: *Annual review of psychology* 66, pp. 223–247.

Smolensky, Paul (1987). "The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn." In: *The Southern Journal of Phulosophy* 26, pp. 137–161. DOI: 10.1111/j.2041-6962.1988.tb00470.x.

Spirtes, Peter, Clark Glymour, and Richard Scheines (1993). "Causation, Prediction, and Search." In: *Springer-Verlag*.

— (2000). *Causation, prediction, and search*. 2nd. Cambridge, MA: MIT Press.

Stenning, Keith and Michiel van Lambalgen (2008). *Human Reasoning and Cognitive Science*. Cambridge, MA: MIT Press.

Sytsma, Justin (2020). "Causation, Responsibility, and Typicality." In: *Review of Philosophy and Psychology* 12.4, pp. 699–719. DOI: 10.1007/s13164-020-00498-2.

Szabolcsi, Anna and Bill Haddican (2004). "Conjunction meets negation: A study in cross-linguistic variation." In: *Journal of Semantics* 21.3, pp. 219–249. DOI: 10.1093/jos/21.3.219.

Tentori, Katya, Vincenzo Crupi, and Selena Russo (2013). "On the determinants of the conjunction fallacy: probability versus inductive confirmation." In: *Journal of Experimental Psychology: General* 142.1, pp. 235–255. DOI: 10.1037/a0028770.

Tishby, Naftali, Fernando C. Pereira, and William Bialek (1999). "The information bottleneck method." In: *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pp. 368–377.

Vasil, Ny, Azzurra Ruggeri, and Tania Lombrozo (Jan. 2022). "When and how children use explanations to guide generalizations." In: *Cognitive Development* 61, p. 101144. ISSN: 0885-2014. DOI: 10.1016/j.cogdev.2021.101144. URL: http://dx.doi.org/10.1016/j.cogdev.2021.101144.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need." In: *Advances in Neural Information Processing Systems*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna Wallach, Robert Fergus, S. V. N. Vishwanathan, and Roman Garnett. Vol. 30. Curran Associates, Inc., pp. 5998–6008.

Walsh, Clare and Philip N. Johnson-Laird (2004). "Co-reference and reasoning." In: *Memory and Cognition* 32, pp. 96–106. DOI: 10.3758/BF03195823.

Weber, Leander, Jim Berend, Moritz Weckbecker, Alexander Binder, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin (2025). *Efficient and Flexible Neural Network Training through Layer-wise Feedback Propagation*. arXiv: 2308.12053 [cs.LG]. URL: https://arxiv.org/abs/2308.12053.

Widrow, Bernard and Marcian E. Hoff (1960). "Adaptive switching circuits." In: *IRE WESCON Convention Record* 4, pp. 96–104.

Woodward, James (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press. DOI: 10.1111/j.1933-1592.2007.00012.x.

— (2006). "Sensitive and insensitive causation." In: *The Philosophical Review* 115.1, pp. 1–50. URL: https://www.jstor.org/stable/20446880.

Yuille, Alan L. and Daniel Kersten (2005). "Vision as Bayesian inference: analysis by synthesis?" In: *Trends in Cognitive Sciences* 10.7, pp. 301–308.

Zaidan, Omar, Jason Eisner, and Christine Piatko (2007). "Using "Annotator Rationales" to Improve Machine Learning for Text Categorization." In: *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. Rochester, NY: Association for Computational Linguistics, pp. 260–267.