รายงานการวิจัย: กรอบแนวคิดทฤษฎีสารสนเทศเพื่อการถ่วงดุลอคติและการสังเคราะห์ข้อมูลเชิงลึกผ่านโครงสร้างลำดับชั้นในปริภูมิแฝง

ความก้าวหน้าของระบบปัญญาประดิษฐ์ในปัจจุบันเผชิญกับความท้าทายที่สำคัญยิ่งในเรื่องของความยุติธรรม (Fairness) และความสามารถในการสังเคราะห์ข้อมูลที่ปราศจากอคติ (Bias) รายงานฉบับนี้วิเคราะห์ถึงแนวทางการสร้างกรอบแนวคิดทางคณิตศาสตร์บนพื้นฐานของทฤษฎีสารสนเทศ (Information Theory) เพื่อบรรลุเงื่อนไขความเท่าเทียมเชิงสารัตถะระหว่างข้อมูลที่มีอคติ (A) และข้อมูลที่ผ่านการปรับสมดุลแล้ว (A') โดยมุ่งเน้นไปที่การรักษาอรรถประโยชน์หลัก (Utility) ขณะที่ลดการรั่วไหลของข้อมูลอ่อนไหว (Sensitive Attributes) ผ่านกลไกการถ่วงน้ำหนักในโครงสร้างลำดับชั้น (Hierarchical Bias Weighting) และการวิเคราะห์เอนโทรปีเชิงธีม (Thematic Entropy) สำหรับข้อมูลเชิงคุณภาพ

รากฐานทฤษฎีสารสนเทศและการนิยามอคติเชิงคณิตศาสตร์

ในบริบทของทฤษฎีสารสนเทศ อคติมิได้เป็นเพียงความลำเอียงเชิงสถิติ แต่คือการละเมิดเงื่อนไขของความเป็นอิสระต่อกันแบบมีเงื่อนไข (Conditional Independence) ระหว่างการทำนายของแบบจำลอง ($\hat{Y}$) และคุณลักษณะที่ต้องได้รับการคุ้มครอง (A) เมื่อพิจารณาจากป้ายกำกับที่แท้จริง (Y) การทำความเข้าใจกลไกนี้จำเป็นต้องอาศัยการกำหนดตัวชี้วัดที่สามารถวัด "ปริมาณ" ของอคติได้อย่างแม่นยำในเชิงคณิตศาสตร์

การวัดอคติผ่านข้อมูลต่างตอบแทนแบบมีเงื่อนไข (Conditional Mutual Information)

การนิยามอคติ B ของแบบจำลอง f บนการกระจายตัวของข้อมูล D สามารถแสดงได้ด้วยสมการข้อมูลต่างตอบแทนแบบมีเงื่อนไข (CMI):
ค่าของ $B > 0$ บ่งชี้ว่าผลลัพธ์จากแบบจำลองยังคงพึ่งพาคุณลักษณะ A (เช่น เชื้อชาติ เพศ หรืออายุ) แม้ว่าจะทราบค่าจริง Y แล้วก็ตาม ในทฤษฎีสารสนเทศเชิงคุณภาพ การวัดนี้สามารถขยายไปสู่ข้อมูลประเภทหมวดหมู่ (Categorical Data) หรือธีมที่เกิดจากการเข้ารหัส (Thematic Coding) โดยการจัดกลุ่มข้อมูลเชิงคุณภาพเข้าสู่พาร์ทิชัน (Partitions) ที่มีความหมายสอดคล้องกับเอนโทรปี
เมื่อ A หมายถึงคุณลักษณะที่ต้องปกป้อง (Protected Attributes) สมการนี้จะกลายเป็นการวัดความยุติธรรมเชิงสถิติ (Statistical Parity) แต่หาก A หมายถึงคุณลักษณะปลอม (Spurious Features) เช่น ฉากหลังของรูปภาพ หรือรูปแบบทางสถิติที่ไม่เกี่ยวข้องกับงานหลัก จะกลายเป็นการวัด "การเรียนรู้ทางลัด" (Shortcut Learning)

ความเท่าเทียมเชิงอรรถประโยชน์ A = A'

เป้าหมายสูงสุดของการถ่วงดุลอคติคือการสร้างตัวแทนข้อมูลใหม่ A' จากข้อมูลเดิม A โดยให้เป็นไปตามเงื่อนไขที่ขัดแย้งกันสองประการ ประการแรกคือการรักษา $I(A'; Y)$ ให้ใกล้เคียงกับ $I(A; Y)$ มากที่สุด เพื่อให้อรรถประโยชน์คงเดิม และประการที่สองคือการทำให้ $I(A'; S) \to 0$ โดยที่ S คือข้อมูลอ่อนไหว สภาวะนี้เรียกว่า "Sufficient Projection" ซึ่งเป็นการฉายข้อมูลเดิมลงในปริภูมิย่อย (Subspace) ที่มีความเป็นอิสระทางสถิติต่อปัจจัยที่เป็นอคติ

ตารางเปรียบเทียบกลไกความเท่าเทียมของอคติรูปแบบต่างๆ:

ประเภทของอคติ	สาเหตุเชิงโครงสร้าง	เงื่อนไขความเท่าเทียมเชิงสารสนเทศ	ผลกระทบต่อประสิทธิภาพ
ความไม่ยุติธรรม (Unfairness)	ความสัมพันธ์เชิงสหสัมพันธ์ระหว่างป้ายกำกับและคุณลักษณะคุ้มครอง	I(\hat{Y}; A | Y) > โดย $ คือ Sensitive Attribute	ความลำเอียงต่อกลุ่มประชากรย่อย
การเรียนรู้ทางลัด (Shortcut Learning)	ความพึงพาคุณลักษณะที่ง่ายต่อการเรียนรู้แต่ไม่มีความสัมพันธ์เชิงสาเหตุ	I(\hat{Y}; S | Y) > โดย $ คือ Spurious Feature	ขาดความทนทานต่อข้อมูลนอกการกระจาย (OOD)
การเปลี่ยนแปลงการกระจายตัว (Distribution Shift)	ความพึงพาสภาพแวดล้อมหรือโดเมนของข้อมูล	I(\hat{Y}; E | Y) > โดย $ คือ Environment Indicator	ประสิทธิภาพลดลงเมื่อนำไปใช้ในบริบทใหม่

กลไกคอขวดสารสนเทศแบบมีเงื่อนไข (Conditional Fairness Bottleneck)

แนวทางที่ได้รับความนิยมสูงสุดในการแก้ปัญหา A = A' คือหลักการคอขวดสารสนเทศ (Information Bottleneck - IB) ซึ่งเดิมถูกใช้เพื่อสกัดข้อมูลที่เกี่ยวข้องมากที่สุดสำหรับการทำนาย ขณะที่บีบอัดข้อมูลที่ไม่จำเป็นออกไป ในงานวิจัยยุคหลัง หลักการนี้ถูกดัดแปลงเป็น "Fair Information Bottleneck" (FairIB)

สมการวัตถุประสงค์และการปรับปรุงความยุติธรรม

สมการวัตถุประสงค์ของ FairIB มุ่งเน้นไปที่การขยายข้อมูลต่างตอบแทนระหว่างตัวแทนข้อมูล (X) และการปฏิสัมพันธ์ที่สังเกตได้ (R) ให้สูงสุด พร้อมกับลดข้อมูลต่างตอบแทนระหว่าง X และคุณลักษณะอ่อนไหว (S) ให้ต่ำสุด:
โดย \beta คือพารามิเตอร์ลากรานจ์ (Lagrange multiplier) ที่ควบคุมขนาดของคอขวด หากข้อมูลมีโครงสร้างที่ซับซ้อน เช่น กราฟ (Graph) หรือเครือข่ายความสัมพันธ์ สารสนเทศที่อ่อนไหวอาจไม่ได้รั่วไหลผ่านโหนดเดียว แต่อาจรั่วไหลผ่านโครงสร้างกราฟย่อย (Sub-graph) ดังนั้นสมการจึงต้องครอบคลุมถึง I(G; S) ซึ่ง G คือตัวแทนของกราฟย่อยด้วย

การประมาณค่าผ่านขอบเขตล่างเชิงแปรผัน (Variational Lower Bounds)

เนื่องจากการคำนวณข้อมูลต่างตอบแทนโดยตรงในปริภูมิที่มีมิติสูงนั้นเป็นปัญหาที่ "คำนวณไม่ได้" (Intractable) จึงจำเป็นต้องอาศัยการประมาณค่าเชิงแปรผัน (Variational Approximation) วิธีการที่สำคัญได้แก่:

1. **Barber & Agakov (IBA):** ใช้เมื่อทราบความหนาแน่นแบบมีเงื่อนไข p(y|x) โดยนำมาเป็นตัวควบคุมขีดความสามารถของตัวแทนข้อมูล
2. **Donsker-Varadhan (IDV):** ให้ขอบเขตล่างที่เข้มงวด แต่อาจมีความแปรปรวนสูงในข้อมูลที่มีความซับซ้อน
3. **Hilbert-Schmidt Independence Criterion (HSIC):** เป็นแนวทางที่ไม่ต้องพึ่งพาข้อสมมติฐานเกี่ยวกับการกระจายตัวของข้อมูล (เช่น ไม่ต้องสมมติว่าเป็น Gaussian) โดยใช้ตัวดำเนินการความแปรปรวนร่วมข้าม (Cross-covariance operator) ในปริภูมิเคอร์เนล

การเลือกใช้ขอบเขตเชิงแปรผันมีผลอย่างมากต่อคุณภาพของตัวแทนข้อมูลที่ได้ ตัวอย่างเช่น การใช้ Deep Amortized Variational Inference สามารถช่วยให้จัดการกับข้อมูลตาราง (Tabular Data) ขนาดเล็กได้อย่างมีประสิทธิภาพ แต่อาจนำไปสู่พฤติกรรมที่แตกต่างกันอย่างมากของระบบหากมีการเลือกโครงสร้างของตัวประมาณค่า (Critic) ที่ไม่เหมาะสม

โครงสร้างลำดับชั้นและการถ่วงน้ำหนักอคติเชิงโหนดลูก

การสังเคราะห์ข้อมูลให้มีความ "ลึกซึ้ง" (Deep/Insightful) สอดคล้องกับโครงสร้างข้อมูลที่มีลักษณะเป็นลำดับชั้น (Hierarchical Structure) ในปริภูมิแฝง (Latent Space) ข้อมูลหนึ่งหน่วยสามารถถูกมองว่าเป็นโหนดแม่ (Parent Node) ที่ประกอบด้วยคุณลักษณะย่อยหรือโหนดลูก (Child Nodes) หลายโหนด

การสกัดและแยกคุณลักษณะ (Disentanglement)

ในกรอบแนวคิดนี้ อคติมักจะแฝงอยู่ในโหนดลูกบางโหนดเท่านั้น หากเราสามารถแยก (Disentangle) คุณลักษณะหลัก (Core Information) ออกจากคุณลักษณะที่เป็นอคติ (Bias-related Features) ได้ เราจะสามารถใช้การถ่วงน้ำหนักเพื่อ "หักล้าง" อคติเหล่านั้นได้
สมการตัวแทนข้อมูลแฝงที่ผ่านการปรับสมดุลแล้วสามารถเขียนได้ในรูป:
โดยที่ Z_{core} คือสารสนเทศหลักที่ไม่เปลี่ยนแปลง Z_k คือคุณลักษณะย่อย (Sub-features) และ \beta_k คือค่าน้ำหนักที่ถูกปรับจูนเพื่อลดผลกระทบจากอคติ ในกรณีของความเสี่ยง (Risk Propagation) น้ำหนักเหล่านี้เรียกว่า Relative Influence (RI) ซึ่งกำหนดว่าสารสนเทศจากโหนดลูกจะส่งผลต่อโหนดแม่มากน้อยเพียงใด

การระบุเอกลักษณ์เชิงสาเหตุในลำดับชั้น (Causal Identifiability)

ปัญหาที่สำคัญของแบบจำลองลำดับชั้นแฝงคือการระบุว่าตัวแปรใดเป็นสาเหตุที่แท้จริง เงื่อนไขการระบุเอกลักษณ์ (Identifiability) สำหรับแบบจำลองเชิงสาเหตุลำดับชั้นแฝงที่ไม่เป็นเส้นตรงกำหนดว่า:

- ตัวแปรแฝงแต่ละตัวต้องมี "โหนดลูกที่บริสุทธิ์" (Pure Children) อย่างน้อย 2 ตัว ซึ่งเป็นตัวแปรที่รับอิทธิพลจากโหนดแม่เพียงโหนดเดียวเท่านั้น
- ต้องไม่มีเส้นทางเชิงสาเหตุโดยตรงระหว่างโหนดพี่น้อง (Siblings) เพื่อป้องกันโครงสร้างสามเหลี่ยมที่ทำให้การคำนวณซับซ้อน

เงื่อนไขเหล่านี้ช่วยให้เราสามารถประมาณค่าตัวแปรแฝงได้จนถึงระดับการแปลงแบบผันกลับได้ (Invertible Transformation) ซึ่งจำเป็นต่อการสร้าง A' ที่ยังคงสาระสำคัญของ A ไว้ได้อย่างครบถ้วน

ทฤษฎีข้อมูลเชิงคุณภาพและการวิเคราะห์เอนโทรปีเชิงธีม

เมื่อข้อมูลเป็นข้อความเชิงพรรณนาหรือธีมเชิงคุณภาพ การวัดความยุติธรรมต้องพึ่งพากลไกการเข้ารหัส (Coding) และการวิเคราะห์โครงสร้างของความหมาย ทฤษฎีสารสนเทศถูกนำมาประยุกต์ใช้เพื่อวัดความไม่แน่นอนในข้อมูลเชิงคุณภาพเหล่านี้

ฟังก์ชันสัญญาณเชิงธีมและเอนโทรปีเชิงธีม H(T)

เอนโทรสนปีเชิงธีมถูกนิยามเพื่อวัดความฟุ้งซ่านหรือ "เสียงรบกวนทางสารสนเทศ" (Information Noise) ในชุดข้อมูลเชิงคุณภาพ หาก p(T_i) คือความน่าจะเป็นของธีมย่อยภายในธีมหลัก T เอนโทรปี H(T) จะถูกคำนวณดังนี้:
เมื่อสัญญาณเชิงธีมมีเสียงรบกวนสูง (อคติเชิงธีม) เอนโทรปีจะเพิ่มขึ้นเนื่องจากมีการให้น้ำหนักกับโทเค็น (Tokens) ที่ไม่มีความหมายเชิงธีม (Thematically Insignificant) มากเกินไป การถ่วงดุลอคติในที่นี้ทำได้โดยการทำ "Universalization" หรือการรวมโทเค็นที่เกี่ยวข้องทางเซมันติก (Semantic) เข้าด้วยกันเพื่อลดเอนโทรปีของพื้นที่ธีมรวม (H(W'))
ตารางความสัมพันธ์ระหว่างความสำคัญของธีมและความไม่แน่นอน:

ระดับความสำคัญของธีม (\theta_k)	ระดับเอนโทรปี (H)	ปริมาณเสียงรบกวน (Z_{noise})	ความหมายเชิงอรรถประโยชน์
สูง	ต่ำ	ต่ำ	ข้อมูลมีความชัดเจนและเป็นกลาง
ต่ำ	สูง	สูง	ข้อมูลมีความกำกวมหรือแฝงอคติมาก
ขีดจำกัดทางทฤษฎี (H_{max})	สูงสุด	สูงสุด	การแบ่งธีมเพิ่มเติมไม่มีประโยชน์เชิงสารสนเทศ

ดัชนีความสำคัญเชิงธีม (Topic Significance Index - TSI)

เพื่อสังเคราะห์ข้อมูลให้ "ลึกซึ้ง" ขึ้น นักวิจัยใช้ TSI ซึ่งเป็นผลคูณของ:

1. **Salience (ความโดดเด่น):** แอมพลิจูดเฉลี่ยของสัญญาณธีมตลอดช่วงเวลา
2. **Purity (ความบริสุทธิ์):** 1 ลบด้วยคะแนนความหลากหลายเชิงความหมาย (S_{div,k})

ธีมที่มี TSI สูงคือธีมที่มีความหมายเฉพาะเจาะจงและมีความสม่ำเสมอ ซึ่งถือเป็น "Core Information" ในเงื่อนไข A = A' ที่เราต้องการรักษาไว้ ขณะที่ธีมที่มี Purity ต่ำอาจสะท้อนถึงการปนเปื้อนของอคติหรือข้อมูลขยะ

การสังเคราะห์มุมมองที่ขัดแย้ง: วิถีแห่งวิภาษวิธี (Dialectics)

ในการสังเคราะห์ข้อมูลให้ยุติธรรมที่สุด จำเป็นต้องนำข้อมูลจากหลายแหล่งที่มีมุมมองแตกต่างกันมาพิจารณา โดยใช้กรอบแนวคิดวิภาษวิธี (Hegelian Dialectic) ซึ่งประกอบด้วย Thesis, Antithesis และ Synthesis

กระบวนการสังเคราะห์สารสนเทศ

- **Thesis (บทตั้ง):** ข้อมูลหรือตัวแทนเดิม A ที่มักแฝงอคติหรือเป็นมุมมองของกลุ่มที่มีอำนาจเหนือกว่า (Dominant Group)
- **Antithesis (บทแย้ง):** ข้อมูลหรือมุมมองจากกลุ่มย่อย (Outgroups) ที่เข้ามาท้าทายอคติใน Thesis เพื่อชี้ให้เห็นข้อบกพร่องเชิงโครงสร้าง
- **Synthesis (บทสังเคราะห์):** การสร้าง A' ที่ก้าวข้ามความขัดแย้ง โดยรวบรวมส่วนที่เป็น "ความจริง" จากทั้งสองฝ่ายและตัดส่วนที่เป็นอคติออก

ในเชิงคณิตศาสตร์ กระบวนการนี้สอดคล้องกับกลไก "Counterfactual Debiasing Inferences" ซึ่งคำนวณผลกระทบทางตรงและทางอ้อมของตัวแปรอคติ แล้วหักล้างมันออกจากผลลัพธ์สุดท้าย เพื่อให้ได้ "Total Indirect Effect" (TIE) ที่สะอาดและลึกซึ้งกว่าเดิม

การประยุกต์ใช้ในข้อมูลเชิงโครงสร้างและพหุรูปแบบ (Multimodal Data)

ความท้าทายของ A = A' ในปัจจุบันขยายตัวไปสู่ข้อมูลที่มีความซับซ้อน เช่น กราฟในระบบการเงิน และแบบจำลองภาษาและภาพ (Vision-Language Models - VLMs)

อคติในกราฟและกลไกคอขวดแฝง

ในโครงสร้างกราฟ อคติไม่ได้เกิดจากคุณสมบัติของโหนดเท่านั้น แต่เกิดจากกลไกการส่งสาร (Message-passing) ที่สะสมอคติจากเพื่อนบ้านที่มีคุณลักษณะคล้ายกัน (Homophily) กรอบงาน GRAFair แก้ปัญหานี้โดยใช้ Variational Graph Autoencoder (VGAE) เพื่อเรียนรู้ตัวแทนข้อมูลที่คงข้อมูลเชิงโครงสร้างไว้ (Structure) แต่ลดข้อมูลต่างตอบแทนกับปัจจัยอ่อนไหว
แนวทาง FairMIB ได้นำเสนอการแยกมุมมองข้อมูลกราฟออกเป็น 3 ส่วน:

1. **Feature View:** คุณลักษณะของโหนด
2. **Structural View:** โทโพโลยีของกราฟที่บริสุทธิ์
3. **Diffusion View:** สารสนเทศในละแวกใกล้เคียงระดับสูง

การใช้ Contrastive Learning เพื่อเพิ่มข้อมูลต่างตอบแทนระหว่างมุมมองที่ต่างกัน (Cross-view MI) ช่วยให้แบบจำลองเรียนรู้ตัวแทนข้อมูลที่มีความคงตัวต่อเสียงรบกวนและอคติเฉพาะจุด

อคติในแบบจำลองพหุรูปแบบ (VLMs)

แบบจำลองอย่าง CLIP มักมีอคติทางเพศและเชื้อชาติเนื่องจากข้อมูลที่ใช้ฝึกฝนมีความไม่สมดุล กลไก Selective Feature Imputation for Debiasing (SFID) ถูกนำเสนอเพื่อแก้ปัญหานี้โดยไม่ต้องฝึกฝนแบบจำลองใหม่ (Training-free) วิธีการนี้ระบุคุณลักษณะที่เป็นอคติโดยใช้เทคนิคอย่าง RandomForest และแทนที่ด้วยตัวแทนที่ไม่มีอคติ (Bias-free Representation) ซึ่งได้มาจากตัวอย่างที่กำกวม (Ambiguous Samples) ทำให้รักษามิติและความหมายเชิงเซมันติกของข้อมูลเดิมไว้ได้อย่างมีประสิทธิภาพ
ตารางเปรียบเทียบวิธีการปรับแก่ในงานด้านต่างๆ:

งาน (Task)	เทคนิคการถ่วงดุลอคติ	เป้าหมายอรรถประโยชน์ (A')	ตัวอย่างอคติที่ถูกกำจัด
การอนุมัติเงินกู้	GNN + Conditional IB	ความสามารถในการชำระคืนตามอาชีพและรายได้	อคติทางเพศที่แฝงในประวัติการชำระ
การแนะนำสินค้า	FairIB + HSIC	Collaborative Signals (ความสนใจที่แท้จริง)	อคติจากประวัติการโต้ตอบที่เบี่ยงเบนตามเชื้อชาติ
การจำแนกภาพ	SFID (Feature Pruning)	ความถูกต้องเชิงโครงสร้างของวัตถุ	ความพึ่งพาสิ่งของที่มักปรากฏร่วมกับเพศเฉพาะ
การวิเคราะห์ข่าว	Thematic Entropy Reduction	ความแม่นยำของหัวข้อหลัก	เสียงรบกวนจากคำศัพท์ที่สะท้อนอคติทางการเมือง

บทสรุป: สู่กรอบแนวคิดความเท่าเทียมที่สมบูรณ์

การสร้างสมการทางคณิตศาสตร์สำหรับการถ่วงดุลอคติภายใต้เงื่อนไข A = A' ไม่ใช่เพียงการลบข้อมูลออก แต่คือการรังสรรค์ปริภูมิแฝงใหม่ที่ข้อมูลหลักถูกเก็บรักษาไว้อย่างสมบูรณ์ภายใต้อิทธิพลของเอนโทรปีที่ถูกจัดระเบียบใหม่ ทฤษฎีสารสนเทศได้มอบเครื่องมืออันทรงพลัง ทั้งในแง่ของตัววัดข้อมูลต่างตอบแทน (MI) และเอนโทรปีเชิงธีม เพื่อช่วยให้นักวิจัยสามารถนำทางท่ามกลางความขัดแย้งของข้อมูลเชิงคุณภาพและเชิงปริมาณได้
กลไกการถ่วงน้ำหนักลำดับชั้น (Hierarchical Weighting) ผ่านโหนดแม่และโหนดลูก ช่วยให้ระบบสามารถเลือกรับเฉพาะสารสนเทศที่มีความสัมพันธ์เชิงสาเหตุ (Causal Relevance) ขณะที่ลดน้ำหนักของคุณลักษณะย่อยที่เป็นแหล่งกำเนิดอคติ การผสานเข้ากับหลักการคอขวดสารสนเทศแบบมีเงื่อนไข (CFB) จะช่วยรับประกันความเสถียรและความทนทานของตัวแทนข้อมูลที่ผ่านการสังเคราะห์แล้ว
ในอนาคต การบูรณาการระหว่างทฤษฎีสารสนเทศและตรรกะเชิงวิภาษวิธีจะช่วยให้การสังเคราะห์ข้อมูลไม่หยุดอยู่แค่ความยุติธรรมเชิงสถิติ แต่ไปถึงความยุติธรรมเชิงสารัตถะที่เข้าใจบริบททางสังคมและประวัติศาสตร์ของข้อมูล นำไปสู่การสร้างปัญญาประดิษฐ์ที่ได้รับความไว้วางใจและมอบข้อมูลเชิงลึกที่ทรงพลังแก่สังคมอย่างแท้จริง

ผลงานที่อ้างอิง

1. When Are Learning Biases Equivalent? A Unifying Framework for Fairness, Robustness, and Distribution Shift - ResearchGate, https://www.researchgate.net/publication/397521755_When_Are_Learning_Biases_Equivalent_A_Unifying_Framework_for_Fairness_Robustness_and_Distribution_Shift 2. Coding Qualitative Data: How To Guide - Thematic, https://getthematic.com/insights/coding-qualitative-data 3. Information Theoretic Framework For Evaluation of Task Level Fairness - GitHub Pages, https://charliezhaoyinpeng.github.io/EAI-KDD22/camera_ready/information.pdf 4. Debiasing with Sufficient Projection: A General Theoretical Framework for Vector Representations - ACL Anthology, https://aclanthology.org/2024.naacl-long.332.pdf 5. Learning Fair Representations for Recommendation via Information Bottleneck Principle - IJCAI, https://www.ijcai.org/proceedings/2024/0273.pdf 6. On variational lower bounds of mutual information - Bayesian Deep Learning, https://bayesiandeeplearning.org/2018/papers/136.pdf 7. On Variational Bounds of Mutual Information - Proceedings of Machine Learning Research, http://proceedings.mlr.press/v97/poole19a/poole19a.pdf 8. Variational Bounds on Mutual Information | Matthew Wiesner,

https://m-wiesner.github.io/Variational-Bounds-on-Mutual-Information/ 9. (PDF) Trustworthy Representation Learning via Information Funnels and Bottlenecks, https://www.researchgate.net/publication/397279389_Trustworthy_Representation_Learning_via_Information_Funnels_and_Bottlenecks 10. A Variational Approach to Privacy and Fairness, https://ppai21.github.io/files/36-paper.pdf 11. Hierarchical Latent Tree Analysis for Topic Detection - Department of Computer Science and Engineering - HKUST, https://www.cse.ust.hk/~lzhang/paper/pspdf/liu-n-ecml14.pdf 12. Evaluating Hierarchical LDA Topic Models for Article Categorization - DiVA portal, https://www.diva-portal.org/smash/get/diva2:1447656/FULLTEXT01.pdf 13. Learning Decomposable and Debiased Representations via Attribute-Centric Information Bottlenecks - arXiv, https://arxiv.org/html/2403.14140v1 14. Bayesian Hierarchical Stacking: Some Models Are (Somewhere) Useful - PMC - NIH, https://pmc.ncbi.nlm.nih.gov/articles/PMC12442486/ 15. A Method for Automatically Eliciting node Weights in a Hierarchical Knowledge-Based Structure for Reasoning with Uncertainty - GRiST Mental Health, https://www.egrist.org/sites/default/files/hegazy-cdb-iaria-09.pdf 16. Identification of Nonlinear Latent Hierarchical Models, https://proceedings.neurips.cc/paper_files/paper/2023/file/065ef23a944b3995de7dd4a3e203d133-Paper-Conference.pdf 17. Integration of Associative Tokens into Thematic Hyperspace: A Method for Determining Semantically Significant Clusters in Dynamic Text Streams - MDPI, https://www.mdpi.com/2504-2289/9/8/197 18. Supervised term-category feature weighting for improved text classification - ResearchGate, https://www.researchgate.net/publication/366562537_Supervised_term-category_feature_weighting_for_improved_text_classification 19. Thesis, Antithesis, Synthesis | Encyclopedia MDPI, https://encyclopedia.pub/entry/32030 20. Storytelling for Oppositionists and Others: A Plea for Narrative - University of Michigan Law School Scholarship Repository, https://repository.law.umich.edu/cgi/viewcontent.cgi?article=3419&context=mlr 21. Persuasive Writing In Three Steps: Thesis, Antithesis, Synthesis - Animalz, https://www.animalz.co/blog/thesis-antithesis-synthesis 22. Understanding Hegel's Dialectical Method: Thesis, Antithesis, Synthesis - PolSci Institute, https://polsci.institute/western-political-thought/hegel-dialectical-method-thesis-antithesis-synthesis/ 23. Thesis, Antithesis, and Synthesis in an Argumentative Essay - The Academic Papers UK, https://www.theacademicpapers.co.uk/blog/2019/06/10/thesis-antithesis-and-synthesis-is-associated-with-which-theory-in-sociology/ 24. US12020480B2 - Counterfactual debiasing inference for compositional action recognition - Google Patents, https://patents.google.com/patent/US12020480B2/en 25. Debiasing Graph Representation Learning Based on Information Bottleneck - IEEE Xplore, https://ieeexplore.ieee.org/iel8/5962385/11022714/10753070.pdf 26. Debiasing Graph Representation Learning based on Information Bottleneck - arXiv, https://arxiv.org/pdf/2409.01367 27. Learning Fair Graph Representations with Multi-view Information Bottleneck | OpenReview, https://openreview.net/forum?id=OhvYHqvlJs 28. LEARNING FAIR GRAPH REPRESENTATIONS WITH MULTI-VIEW INFORMATION BOTTLENECK - OpenReview, https://openreview.net/pdf/92e7f2de4f3ecdc360e46154f66ac9fd8c7923b0.pdf 29. Learning Fair Graph Representations with Multi-view Information Bottleneck - arXiv, https://arxiv.org/html/2510.25096v1 30. A Unified Debiasing Approach for Vision-Language Models across Modalities and Tasks - NIPS papers, https://proceedings.neurips.cc/paper_files/paper/2024/file/254404d551f6ce17bb7407b4d6b3c87b-Paper-Conference.pdf 31. An Information Theoretic Approach to Reducing Algorithmic Bias for Machine Learning, https://www.researchgate.net/publication/360647641_An_Information_Theoretic_Approach_to_

Reducing_Algorithmic_Bias_for_Machine_Learning 32. Rethinking Media Literacy: A New Ecosystem Model for Information Integrity - World Economic Forum: Publications, https://reports.weforum.org/docs/WEF_Rethinking_Media_Literacy_2025.pdf 33. Involvement of People With Dementia in the Development of Technology-Based Interventions: Narrative Synthesis Review and Best Practice Guidelines - NIH, https://pmc.ncbi.nlm.nih.gov/articles/PMC7746489/