Folk Theories, Machine Learning, and XAI

(an early-stage investigation)

Can what the average, non-expert person believes to be true about machine learning (their folk theories or mental models) facilitate more effective explanations (XAI)?

Folk theories are "the mental representations that humans use to structure experience". (Gelman & Legare, 2011)

Folk theories "need not be technically accurate (and usually are not), but they must be functional." (Norman 1983)

As a result, folk theories are explanatory. They can be leveraged to align XAI strategies with user beliefs)

Part of the DARPA definition of XAI is to "enable users to **understand**, appropriately **trust**, and effectively **manage**" machine learning systems. (Launchbury 2017)

This is what people say (their folk theories) about recommender systems

Operational (Process): "control system", "database", "code", "taught", "statistical", "feedback"

Abstract (Personification): "buddy", "hacker", "intense", "annoying", "greedy", "always knows what I want"

Power (Control): "submission", "resistance", "balance", "capitulation"

Contrarian (Disbelief): "random", "human controlled"

(From published user studies)

Can we align user folk theories with DARPA's XAI user objectives to identify focused XAI strategies?

For example:

If the user's dominant folk theory is …	The key XAI objective is …	Primary XAI user focus should be …
Operational	Understanding ("need" to know; "want" to know)	Accuracy
Abstract	Trust (experiential; dynamic)	Predictability
Power	Management (objectives; skillscontextual)	Competence

Estimates suggest 12% of the lay population are contrarians. (Eslami et al. 2017) What to do about t he Contrarians?

Michael Ridley: PhD Candidate, Western University;
Postgraduate Affiliate, Vector Institute; mridley@uoguelph.ca