One if by land, two if by sea, three if by four seas, and more to come — values of perception, prediction, communication, and common sense in decision making

Aolin Xu

Abstract

This work is about rigorously defining the values of perception, prediction, communication, and common sense in decision making. The defined quantities are decision-theoretic, but have information-theoretic analogues, e.g., they share some simple but key mathematical properties with Shannon entropy and mutual information, and can reduce to these quantities in particular settings. One interesting observation is that, the value of perception without prediction can be negative, while the value of perception together with prediction and the value of prediction alone are always nonnegative. The defined quantities suggest answers to practical questions arising in the design of autonomous decision-making systems. Example questions include: Do we need to observe and predict the behavior of a particular agent? How important is it? What is the best order to observe and predict the agents? The defined quantities may also provide insights to cognitive science and neural science, toward the understanding of how natural decision makers make use of information gained from different sources and operations.

1 Introduction

This work is motivated by the following questions arising in decision-making problems.

- What is the value of observing the world when making decisions, compared to not observing the world at all?
- When we can only observe part of the world but can predict the rest, what is the value of predicting the unobserved part of the world?
- When there are multiple agents involved in decision making, how important is it to observe and predict the behavior of a particular agent? and
- What is the best order to perform perception and prediction on these agents?

Answering the first two questions can help us quantify the importance of the perception and prediction subsystems in a decision-making system. The latter two questions are of particular interest when the limited computational resource is a concern in the design of the decision-making system. In this work, we attempt to provide a rigorous path to formalize the above questions and derive answers to them. We define the values of key operations for decision making, including perception and prediction, and extending to communication and

The author is with Honda Research Institute US. Email: xuaolin@gmail.com

NeurIPS 2023 Workshop: Information-Theoretic Principles in Cognitive Systems (InfoCog)

common sense; we also extend these definitions to decision-making problems with multiple agents. We analyze the mathematical properties of these values and draw connections to information-theoretic measures. The defined quantities may also provide insights to cognitive and neural sciences, toward the understanding of how natural decision makers make use of information gained from different sources and operations. Related works are discussed in Appendix A.

2 Definitions of values of different operations

2.1 A generic decision-making problem

This study concerns a generic decision-making problem, where a decision maker need to take an action u from action space U in order to minimize the expectation $\mathbb{E}[\ell(X, Y, u)]$ of a loss function $\ell : \mathsf{X} \times \mathsf{Y} \times \mathsf{U} \to \mathbb{R}$, where X is a random observable state, Y is a random unobservable state, and the expectation is taken with respect to the joint distribution $P_{X,Y}$. X being observable means that the action can be taken as a function of X through a policy $\psi : \mathsf{X} \to \mathsf{U}$, such that $U = \psi(X)$. This problem encompasses several types of problems, e.g.

- One-step or open-loop stochastic control, where the loss depends on both the observed and unobserved states. This includes the settings where the loss only depends on (X, u) or (Y, u) as special cases.
- Bayesian statistical estimation, including both regression and classification, where Y need to be estimated based on X. An advantage of considering both X and Y in the loss function is that it makes the loss contextual, which can better model the situations where the quality of an estimate depends on its accuracy as well as the context it is made in.
- Bayesian game play with an agent, where X includes the ego player's own state and the observable state of the agent, Y represents the unknown action to be taken by the agent that only depends on X, and the loss depends on both players' states and actions.

For this decision-making problem, how well the decision can be made depends on what information or knowledge the decision maker uses when coming up with the action. Possible knowledge that can be used includes the marginal distributions P_X and P_Y, the joint distribution $P_{X,Y}$, the observable state X, and even the unobservable state Y. There are also different ways of using $P_{X,Y}$, depending on whether it is used together with X or not. Gaining and making use of these knowledge usually requires different operations, such as common sense inquiry, perception, prediction, and even communication. In general, the more knowledge is gained, the better decision can be made. In what follows, we present a rigorous way to define the operations for gaining different knowledge, and quantify the values of these operations when the gained knowledge is used for decision making.

2.2 Value of common sense

When neither X nor Y can be used for decision making, knowing their joint distribution can still be helpful as it encodes their statistical dependency, which is usually gained from common sense. We can thus quantify the value of common sense as the reduction of the minimum achievable expected loss, a.k.a. the risk, when using $P_{X,Y}$ in decision making, compared with when ignoring the statistical dependency between X and Y by using only the product of their marginals $P_X P_Y$. Formally, this value can be defined as

$$\mathbb{E}[\ell(X, Y, \bar{u})] - \mathbb{E}[\ell(X, Y, u^*)] \tag{1}$$

where $\bar{u} \triangleq \arg\min_{u \in \mathsf{U}} \mathbb{E}[\ell(X, \bar{Y}, u)]$ with $P_{X,\bar{Y}} = P_X P_Y$ is the best action to take when X and Y are treated to be independent from each other; and $u^* \triangleq \arg\min_{u \in \mathsf{U}} \mathbb{E}[\ell(X, Y, u)]$ is the best action to take when using the true joint distribution of X and Y. From the definition of u^*, we know that the value of common sense is always nonnegative.

2.3 Value of perception

The observable state X can be used in decision making only when it is perceived. In a narrow sense, perception is the operation to get X ready for use by a policy, while not taking the dependency of Y on X into account. We can define the value of this narrow sense of perception as the reduction of risk when using X in decision making without predicting Y based on X, compared with only using $P_{X,Y}$ in decision making. Formally, it can be

$$\mathbb{E}[\ell(X, Y, u^*)] - \mathbb{E}[\ell(X, Y, \bar{\psi}(X))] \tag{2}$$

where $\bar{\psi} \triangleq \arg\min_{\psi:\mathsf{X}\to\mathsf{U}} \mathbb{E}[\ell(X, \bar{Y}, \psi(X))]$ is the best policy when using X but treating Y to be independent of X in decision making. Given $X = x$, $\bar{\psi}(x)$ can be realized as $\bar{\psi}(x) = \arg\min_{u\in\mathsf{U}} \mathbb{E}[\ell(x, Y, u)]$. In other words, $\bar{\psi}$ only takes X and the marginal distribution P_Y into account when coming up with the best action to take, thus does not predict Y based on X. There is no definite order between $\mathbb{E}[\ell(X, Y, u^*)]$ and $\mathbb{E}[\ell(X, Y, \bar{\psi}(X))]$, meaning that the value of perception defined above could be negative for certain $P_{X,Y}$ and ℓ. One such example is given in Appendix B. An intuitive example is that, it may be safer for a driver to pull up the vehicle when they cannot clearly see the surrounding objects, than driving with limited consciousness while only reacting to immediately visible objects but ignoring the interactions between them and the invisible ones. The full potential of perception needs to be realized by the additional operation of prediction, as discussed in the next two subsections.

2.4 Value of prediction

The value of prediction can be defined as the reduction of risk when making use of X as well as predicting Y based on it in decision making, compared with just using X and treating Y as independent of X. Formally, it can be defined as

$$\mathbb{E}[\ell(X, Y, \bar{\psi}(X))] - \mathbb{E}[\ell(X, Y, \psi^*(X))] \tag{3}$$

where $\psi^* \triangleq \arg\min_{\psi:\mathsf{X}\to\mathsf{U}} \mathbb{E}[\ell(X, Y, \psi(X))]$ is the best policy when using X and the true joint distribution $P_{X,Y}$ in decision making. Given $X = x$, $\psi^*(x)$ can be realized as $\psi^*(x) = \arg\min_{u\in\mathsf{U}} \mathbb{E}[\ell(x, Y, u)|X = x]$. In other words, ψ^* takes both X and the posterior $P_{Y|X}$ into account when coming up with the best action to take. Computationally, the operation of predicting Y based on X means the evaluation of the posterior $P_{Y|X=x}$ for each given x. From the definition of ψ^*, we know that the value of prediction is always nonnegative.

2.5 Value of perception together with prediction

We can also define the value of perception together with prediction as the reduction of risk when using X and predicting Y in decision making, compared with only using $P_{X,Y}$:

$$\mathbb{E}[\ell(X, Y, u^*)] - \mathbb{E}[\ell(X, Y, \psi^*(X))] \tag{4}$$

which simply equals the sum of the value of perception and the value of prediction defined above. From the definition of ψ^*, we know that the value of perception together with prediction is always nonnegative.

2.6 Value of communication

When the decision maker not only uses observable X, but also can access Y through some form of communication, the achievable risk in the decision making can be further reduced. This reduction can be defined as the value of communication:

$$\mathbb{E}[\ell(X, Y, \psi^*(X))] - \mathbb{E}[\ell(X, Y, \psi^{**}(X, Y)] \tag{5}$$

where $\psi^{**} \triangleq \arg\min_{\psi:\mathsf{X}\times\mathsf{Y}\to\mathsf{U}} \mathbb{E}[\ell(X, Y, \psi(X, Y)]$ is the best communication-enabled policy that uses both X and Y and their joint distribution in decision making. Since a communication-enabled policy defined as $\psi(x, y) = \psi^*(x)$ that simply ignores Y can achieve $\mathbb{E}[\ell(X, Y, \psi^*(X))]$, we know that the value of communication is always nonnegative.

3 Information-theoretic realizations and interpretations

The values of different operations defined above have information-theoretic realizations as commonly seen information measures. These information measures in turn provide information-theoretic interpretations and analogues of the defined values. Consider the case where X and Y are jointly distributed discrete random variables, the action space is the set of joint distributions on $\mathsf{X} \times \mathsf{Y}$, and the loss function is $\ell(x, y, u) = -\log u(x, y)$, that is, the negative log-likelihood of (x, y) under distribution u. In this case, the best actions and policies defined above become $\bar{u} = P_X P_Y$, $u^* = P_{X,Y}$, $\bar{\psi}(x) = \delta_x P_Y$, $\psi^*(x) = \delta_x P_{Y|X=x}$, $\psi^{**}(x, y) = \delta_x \delta_y$, where δ_x and δ_y are one-hot distributions at x and y respectively. The various values defined above become: value of common sense $H(X) + H(Y) - H(X, Y) = I(X; Y)$, value of perception $H(X, Y) - H(Y) = H(X|Y)$, value of prediction $H(Y) - H(Y|X) = I(X; Y)$, value of perception together with prediction $H(X, Y) - H(Y|X) = H(X)$, and value of communication $H(Y|X) - 0 = H(Y|X)$. It is interesting to note that in this information-theoretic setting, the values of common sense and prediction both are the mutual information between X and Y. It is also notable that the uncertainty of X can be fully removed only when both observing X and predicting Y based on X, as

$$H(X, Y) - H(Y|X) = H(X|Y) + I(X; Y) = H(X). \tag{6}$$

The defined quantities and their information-theoretic realizations are illustrated in Fig. 1. The connections between the defined values and the information measures are further studied through their functional properties in Appendix C.

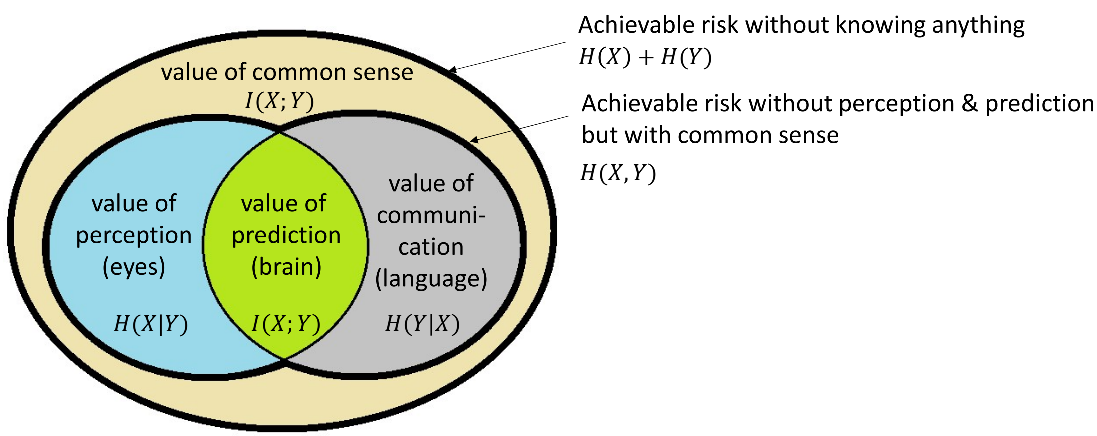

Figure 1: Values of common sense, perception, prediction, and communications, as well as their information-theoretic analogues.

4 Extension to multi-agent settings and applications

We can extend the definitions of the values of different operations to decision-making problems involving multiple agents. The extended results can provide ground truths of agent importance and the best order of perception and prediction of the agents in such problems.

Consider the problem where the decision maker needs to take an action $u \in \mathsf{U}$ to minimize the expected loss $\mathbb{E}[\ell(X^n, Y^n, u)]$ influenced by n agents. Each agent has an observable state X_i which could be their status, and an unobservable state Y_i which could be their action to take, for $i = 1, \dots, n$. There is a joint distribution P_{X^n,Y^n} of $X^n = (X_1, \dots, X_n)$ and $Y^n = (Y_1, \dots, Y_n)$, known to the decision maker. The action can be taken based on P_{X^n,Y^n} and X^n. Using the concepts developed in the previous sections, we may gauge the importance of an agent by examining the value of observing and making prediction on that agent. For the ith agent, this value can be defined as

$$\mathbb{E}[\ell(X^n, Y^n, u^*)] - \mathbb{E}[\ell(X^n, Y^n, \psi_i^*(X_i))] \tag{7}$$

where we redefine u^* to be $u^* = \arg\min_{u \in \mathsf{U}} \mathbb{E}[\ell(X^n, Y^n, u)]$, and define $\psi_i^* = \arg\min_{\psi: \mathsf{X} \to \mathsf{U}} \mathbb{E}[\ell(X^n, Y^n, \psi(X_i))]$. This value reflects the importance of taking the ith agent into account for decision making without or before taking any other agent into account. Further discussions on the definitions of these values and the order of importance of the agents are continued in Appendix D.

Acknowledgment

The author would like to thank Behzad Dariush, Teruhisa Misu, and Enna Sachdeva for helpful discussions. The title of the paper follows the title of a talk given by Prof. Robert Gallager on the occasion of Prof. Bruce Hajek's sixtieth birthday in Oct. 2015 at CSL UIUC.

Appendix

A Related works

This work can be viewed as an extension of (1) on the generalized definition of entropy and mutual information in decision making, which was subsequently studied in (2) in game theoretic setting. The value of information is also discussed in economics setting in (3). In those earlier works, the "value of information" can be viewed as the special case of the value of perception together with prediction as defined in this work. In this work, we specifically distinguish between perception and prediction, and extent the discussion to cover common sense and communication as well. The decision-making problem considered here is also more general, due to the more general form of the loss function we use. The loss function used here follows the one considered in (4; 5). To the author's knowledge, there is no dedicated study in the literature on the values of the diverse operations considered in this work yet. The nonnegativity of the most values considered here also generalizes the data processing inequality of Bayes risk in (6).

The identification of importance of agents in autonomous driving has been a trending topic in recent years. Example works include (7; 8; 9; 10). Those works are mostly based on machine learning methods which require human labels on the importance of the agents, or based on rules from human knowledge. The results in this work can be used to provide ground truth labels of agent importance to be used for the learning-based methods.

Recently, common sense reasoning in perception, prediction, and planning becomes a trending topic in AI research (11). Gaining a principled understanding of the importance of common sense helps to promote the related research in a more scientific manner.

B An example where value of perception is negative

One such example is where $\mathsf{X} = \mathsf{Y} = \mathsf{U} = \{0, 1\}$, the loss function is

$(x,\, y,\, u)$	000	001	010	011	100	101	110	111
ℓ	1	0	0	1	0	0	1	0

and the joint distribution $P_{X,Y}$ and the conditional distribution $P_{Y|X}$ are

$(x,\, y)$	00	01	10	11
$P_{X,Y}$	0.1	0.05	0.15	0.7

$(x,\, y)$	00	01	10	11
$P_{Y\|X}$	0.67	0.33	0.18	0.82

For this example, $u^* = 1$ and $\mathbb{E}[\ell(X, Y, u^*)] = 0.05$, while $\bar{\psi}(0) = 0$, $\bar{\psi}(1) = 1$, and $\mathbb{E}[\ell(X, Y, \bar{\psi}(X))] = 0.1$, resulting in a negative value of perception.

C Functional properties of defined values

We have seen that the defined values are always nonnegative except for the value of perception without prediction. In this section we further study some functional properties of the defined values, mainly concavity and convexity in the underlying distributions. These properties are inherited as key properties by their information-theoretic realizations. This further reveals the intimate connection between the defined values and the information measures.

The results are built on four lemmas. The first three are about properties of the achievable risks, which are useful in proving properties of the value of perception together with prediction and the value of communication stated in Propositions 1 and 2.

Lemma C1. *$\mathbb{E}[\ell(X, Y, u^*)]$ is concave in the joint distribution $P_{X,Y}$.*

Lemma C2. *$\mathbb{E}[\ell(X, Y, \psi^*(X))]$ has the following properties:*

1. *For a fixed P_X, it is concave in $P_{Y|X}$.*

2. *For a fixed $P_{Y|X}$, it is linear in P_X.*

Lemma C3. *$\mathbb{E}[\ell(X, Y, \psi^{**}(X, Y))]$ is linear in the joint distribution $P_{X,Y}$.*

The next lemma is useful in proving the properties of the values of common sense and prediction stated in Proposition 3.

Lemma C4. *Given a loss function $\ell : \mathsf{Z} \times \mathsf{U} \to \mathbb{R}$ and two distributions P and Q on Z, a generalized decision-theoretic divergence between P and Q with respect to ℓ can be defined as*

$$D_\ell(P, Q) = \mathbb{E}_P[\ell(Z, u_Q)] - \mathbb{E}_P[\ell(Z, u_P)] \tag{8}$$

where $u_P = \arg\min_{u \in \mathsf{U}} \mathbb{E}_P[\ell(Z, u)]$ and $u_Q = \arg\min_{u \in \mathsf{U}} \mathbb{E}_Q[\ell(Z, u)]$. This divergence is convex in P when Q is fixed.

Now we can state the major functional properties in the following three propositions.

Proposition 1. *The value of perception together with prediction is concave in P_X for a fixed $P_{Y|X}$.*

This result echoes the information-theoretic realization of the value of perception together with prediction as $H(X)$, which is concave in P_X. This value can thus be seen as a generalized form of entropy of X, implying that the uncertainty about X in decision making can be fully resolved by observing it and making prediction on Y based on it.

Proposition 2. *The value of communication have the following properties:*

1. *For a fixed P_X, it is concave in $P_{Y|X}$.*

2. *For a fixed $P_{Y|X}$, it is linear in P_X.*

This result echoes the information-theoretic realization of the value of communication as $H(Y|X)$, which has the same properties. It shows that this value can be seen as a generalized form of conditional entropy of Y given X, implying that the residue uncertainty of Y after predicting it based on X in decision making can be fully resolved by communication.

Proposition 3. *For a fixed pair of P_X and P_Y, and the set of probability transition kernels $\mathcal{M} = \{P_{Y|X} : P_{Y|X} \circ P_X = P_Y\}$ that couple the pair, the value of common sense and the value of prediction both are convex in $P_{Y|X}$ on $\mathcal{M}$.*

This result echoes the information-theoretic realization of the values of common sense and prediction as $I(X;Y)$, which has a stronger property that it is convex in $P_{Y|X}$ as long as P_X is fixed. It shows that these two values both can be seen as a generalized form of mutual information between X and Y, implying that the uncertainty of Y in decision making can be partially resolved by using its statistical dependency on X.

D Further discussions on decision making with multiple agents

Another way of defining the value or importance of the ith agent would be the reduction of risk by observing and making prediction based on this agent while the other agents are already taken into account, that is

$$\mathbb{E}[\ell(X^n, Y^n, \psi^{*n\setminus i}(X^{n\setminus i}))] - \mathbb{E}[\ell(X^n, Y^n, \psi^{*n}(X^n))] \tag{9}$$

where

$$\psi^{*n\setminus i} = \underset{\psi:\mathsf{X}^{n-1}\to\mathsf{U}}{\arg\min}\, \mathbb{E}[\ell(X^n, Y^n, \psi(X^{n\setminus i}))], \tag{10}$$

and

$$\psi^{*n} = \underset{\psi:\mathsf{X}^{n}\to\mathsf{U}}{\arg\min}\, \mathbb{E}[\ell(X^n, Y^n, \psi(X^n))] \tag{11}$$

are the best policies when taking $X^{n\setminus i} \triangleq (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ and X^n respectively into account for decision making.

As both definitions are viable for agent importance, we see that the importance of an agent actually depends on its order in perception and prediction. In other words, it depends on which other agents have already been taken into account for decision making before observing and making predictions based on that agent. This motivates us to continue this chain of thought to answer the second question.

The best order $(i_1, \ldots, i_n)$ to run perception and prediction on agents can be determined as

$$i_1 = \operatorname*{arg\,min}_{i \in [n]} \mathbb{E}[\ell(X^n, Y^n, \psi_i^*(X_i))] \tag{12}$$

and for $k = 2, \ldots, n$

$$i_k = \operatorname*{arg\,min}_{i \in [n] \setminus \{i_1, \ldots, i_{k-1}\}} \mathbb{E}[\ell(X^n, Y^n, \psi^*_{i^{k-1}, i}(X_{i_1}, \ldots, X_{i_{k-1}}, X_i))] \tag{13}$$

where

$$\psi^*_{i^{k-1}, i} = \operatorname*{arg\,min}_{\psi : \mathsf{X}^k \to \mathsf{U}} \mathbb{E}[\ell(X^n, Y^n, \psi(X_{i_1}, \ldots, X_{i_{k-1}}, X_i))]. \tag{14}$$

Equivalently, i_1 maximizes among $i \in [n]$ the value of the first agent to take into account

$$\mathbb{E}[\ell(X^n, Y^n, u^*)] - \mathbb{E}[\ell(X^n, Y^n, \psi_i^*(X_i))] \tag{15}$$

while i_k maximizes among $i \in [n] \setminus \{i_1, \ldots, i_{k-1}\}$ the value of the kth agent to take into account.

References

[1] M. H. DeGroot, "Uncertainty, Information, and Sequential Experiments," *The Annals of Mathematical Statistics*, vol. 33, no. 2, pp. 404 – 419, 1962.

[2] P. D. Grünwald and A. P. Dawid, "Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory," *The Annals of Statistics*, vol. 32, no. 4, pp. 1367 – 1433, 2004.

[3] R. A. Howard, "Information value theory," *IEEE Transactions on Systems Science and Cybernetics*, vol. 2, no. 1, pp. 22–26, 1966.

[4] A. Xu, "Dynamic inference," *arXiv preprint, arXiv:2111.14746*, 2021.

[5] A. Xu and P. Guan, "Bayesian learning for dynamic inference," *arXiv preprint, arXiv:2301.00032*, 2022.

[6] A. Xu and M. Raginsky, "Minimum excess risk in Bayesian learning," *IEEE Transactions on Information Theory*, vol. 68, no. 12, pp. 7935–7955, 2022.

[7] K. S. Refaat, K. Ding, N. Ponomareva, and S. Ross, "Agent prioritization for autonomous navigation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[8] M. Gao, A. Tawari, and S. Martin, "Goal-oriented object importance estimation in on-road driving videos," in *International Conference on Robotics and Automation (ICRA)*, 2019.

[9] Z. Zhang, A. Tawari, S. Martin, and D. J. Crandall, "Interaction graphs for object importance estimation in on-road driving videos," *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[10] J. Li, H. Gang, H. Ma, M. Tomizuka, and C. Choi, "Important object identification with semi-supervised learning for autonomous driving," in *International Conference on Robotics and Automation (ICRA)*, 2022.

[11] Y. Hong, B. Wu, Z. Chen, Q. Zhou, and C. Gan, "Workshop on machine visual common sense: Perception, prediction, planning," *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.